UNIVERSITÀ DEGLI STUDI DI MACERATA

CORSO DI DOTTORATO DI RICERCA IN

UMANESIMO E TECNOLOGIE

CICLO XXXV

TITOLO DELLA TESI

Lo sviluppo di un framework di estrazione automatica di metadati per la conservazione dei prodotti della ricerca

SUPERVISORE DI TESI
Chiar.ma Prof. Anna Rovella

DOTTORANDO
Dott. Armando Bartucci

COORDINATORE
Chiar.mo Prof. Roberto Lambertini

ANNO 2023

Sommario

Introduzione

La ricerca svolge un ruolo importante nel progresso culturale e scientifico di una società e nello sviluppo economico dei suoi territori. Ciò malgrado i risultati, i dati, i documenti che caratterizzano i processi di produzione della ricerca e che sono alla base della possibilità di un riuso fruttuoso della stessa rimangono troppo spesso inaccessibili se non al ristretto manipolo degli autori. Già da diversi anni il movimento Open Access promuove la possibilità di accesso aperto alle pubblicazioni scientifiche, nella piena consapevolezza che tale aspetto costituisca solo la punta di un iceberg tanto imponente quanto profondo che può essere combattuto solo con un approccio di Open Science. "Secondo i principi dell'Open Science il sapere prodotto con risorse pubbliche è una ricchezza da valorizzare invece che da chiudere. Qualsiasi prodotto della ricerca pubblica deve essere disponibile gratuitamente e deve poter essere liberamente riutilizzabile per aumentare l'impatto del lavoro scientifico" .

Partendo da tali considerazioni il presente lavoro di tesi si orienta verso una lettura dei prodotti di ricerca in una duplice chiave:

- la natura di oggetti digitali dei prodotti della ricerca e il loro ruolo documentale negli archivi digitali degli enti di ricerca;
- i fabbisogni informativi e le esigenze di conservazione e tutela di tale patrimonio informativo.

In particolare la ricerca condotta per la presente tesi di dottorato, muovendo dall'idea della conservazione digitale dei prodotti della ricerca mira a definire un framework di estrazione automatica di metadati in grado di ottimizzare l'accesso a tali importanti fonti informative e di conoscenza.

Nel 2018 la Commissione europea ha pubblicato una serie di raccomandazioni sull'accesso all'informazione scientifica rendendo obbligatori progressivamente strumenti e metodi di governo della conoscenza sviluppata durante la gestione dei progetti di ricerca finanziati attraverso i programmi quadro e le altre iniziative UE. Si tratta di un primo importante passo che, tuttavia, richiede ancora un lungo e faticoso lavoro prima di giungere a traguardi davvero significativi.

Nonostante l'impegno dei ricercatori i dati, le informazioni e i documenti della ricerca non entrano ancor oggi a far parte dei sistemi di conservazione se non in una misura parziale e spesso non significativa.

Obiettivo del lavoro di tesi è allora anche quello di un'analisi e una riflessione sulla reale possibilità di conservazione di tali risorse e sull'occassione di ridisegnare i confini della scienza aperta in una logica di salvaguardia e di affidabilità dei prodotti. In tale prospettiva il dibattito si sposta sulla tematica dell'accesso agli oggetti conservati come risultato di un coerente processo di rappresentazione dei metadati. Come è noto una metadatazione di qualità richiede, generalmente, l'investimento di considerevoli somme di denaro e un importante dispendio di tempo. L' attuale sviluppo di tecniche di Machine e Deep Learning consente di implementare sistemi di estrazione di

conoscenza sempre più precisi e performanti. Il dominio della scienza del libro e del documento è senza dubbio un ottimo banco di prova per l'applicazione di tool e di estrazione automatica di metadati. Il framework teorizzato nel presente lavoro prova, quindi, ad automatizzare il processo di estrazione dei metadati anche in una logica di escosistemi digitali in cui l'interoperabilità diventa uno dei principali asset. La comunità scientifica si riconfigura come uno dei possibili stakeholder di una rete in cui decisori, imprenditori e enti di ricerca operano sinergicamente alla costruzione di innovazione e sviluppo per il benessere dei cittadini.

Al termine del lavoro di tesi una riflessione sui risultati ottenuti appare necessaria. In primis occorre evidenziare la complessità del dominio relativo ai progetti di ricerca.

Un'ulteriore criticità è rappresentata dall'ambiente digitale in cui si trovano i prodotti della ricerca che se da un alto agevola i processi di comunicazione dall'altro richiama l'applicazione di modelli e metodologie per la rappresentazione delle risorse e per la garanzia di un loro recupero nel tempo senza comprometterne il valore probatorio.

Vi è poi una difficoltà riconducibile alla scarsa cultura dei ricercatori verso i processi di gestione dei dati e delle informazioni in un contesto che, troppo spesso, deve fare i conti con tempi sempre più stretti e elevati parametri quali/quantitativi da rispettare.

Altro fattore frenante è la continua pressione esercitata da editori e grandi player dell'informazione che contrastano con forza la possibilità di realizzazione della scienza aperta.

Senza dubbio insufficiente risulta, anche, l'attenzione da parte delle governance degli enti di ricerca e infine, non meno importante è la ridotta domanda di accesso da parte dei decisori e degli attori del sistema produttivo.

Le criticità elencate disegnano un quadro articolato e complesso in cui gli elementi individuati nella tesi possono solo rappresentare approcci e soluzioni parziali che possono diventare abilitanti in presenza di una logica di sinergia e di stakeholder engagement.

Per tale ragione il modello di framework definito nella tesi prova ad ancorarsi saldamente al dominio, al contesto di produzione della ricerca per selezionare tecnologie, metodologie e processi operativi adeguati alle specifiche dell'ambiente.

Nei prossimi mesi il modello sarà sperimentato su alcune community al fine di valutarne l'efficacia e/o la necessità di modifiche o integrazioni.

Il lavoro di tesi è organizzato in 2 capitoli:

- il capitolo 1 traccia lo stato dell'arte relativamente ai prodotti della ricerca, ai metadati e alla conservazione digitale.
- il capitolo 2 partendo di modelli di conservazione e dalle metodologie e tecniche di estrazione automatica di metadati arriva a definire la struttura concettuale del framework di estrazione e l'architettura tecnologica dello stesso.

Al termine del lavoro di tesi una riflessione sui risultati ottenuti appare necessaria. In primis occorre evidenziare la complessità del dominio relativo ai progetti di ricerca.

Un'ulteriore criticità è rappresentata dall'ambiente digitale in cui si trovano i prodotti della ricerca che se da un alto agevola i processi di comunicazione dall'altro richiama l'applicazione di modelli e metodologie per la rappresentazione delle risorse e per la garanzia di un loro recupero nel tempo senza comprometterne il valore probatorio.
Vi è poi una difficoltà riconducibile alla scarsa cultura dei ricercatori verso i processi di gestione dei dati e delle informazioni in un contesto che, troppo spesso, deve fare i conti con tempi sempre più stretti e elevati parametri quali/quantitativi da rispettare.
Altro fattore frenante è la continua pressione esercitata da editori e grandi player dell'informazione che contrastano con forza la possibilità di realizzazione della scienza aperta.
Senza dubbio insufficiente risulta, anche, l'attenzione da parte delle governance degli enti di ricerca e infine, non meno importante è la ridotta domanda di accesso da parte dei decisori e degli attori del sistema produttivo.
Le criticità elencate disegnano un quadro articolato e complesso in cui gli elementi individuati nella tesi possono solo rappresentare approcci e soluzioni parziali che possono diventare abilitanti in presenza di una logica di sinergia e di stakeholder engagement.
Per tale ragione il modello di framework definito nella tesi prova ad ancorarsi saldamente al dominio, al contesto di produzione della ricerca per selezionare tecnologie, metodologie e processi operativi adeguati alle specifiche dell'ambiente.
Nei prossimi mesi il modello sarà sperimentato su alcune community al fine di valutarne l'efficacia e/o la necessità di modifiche o integrazioni.

1. Scenario e stato dell'arte

1.1. Prodotti della ricerca

«*Open Science is just good science*! »
(Jon Tennant, 2018)

I prodotti della ricerca sono in costante incremento. Le stime fornite da *DBLP computer science bibliography* (*Fig. 1*) indicano che il numero nel database *DBLP* è aumentato da 10.029 nel 1995 a 2.177.485 nel marzo 2021[1].

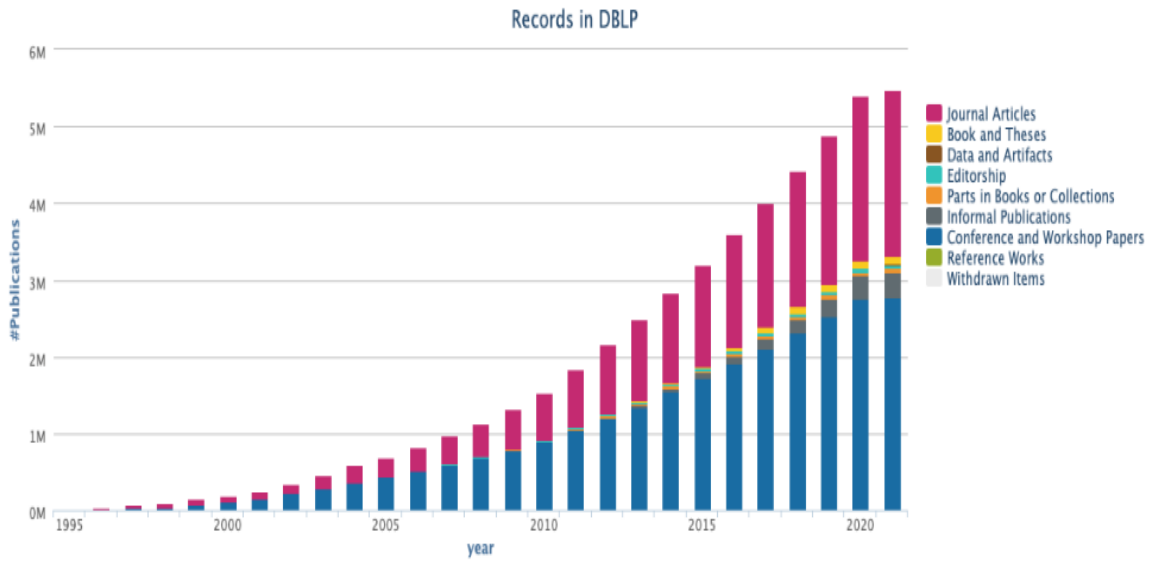

Figura 1. Il diagramma mostra il numero totale delle diverse tipologie di pubblicazioni scientifiche (https://dblp.uni-trier.de/statistics/recordsindblp.html)

L'archivio di *PubMed Central* (PMC)[2] attualmente conserva più di 6 milioni di record full-text inerenti la letteratura biomedica rispetto ai 2 milioni del 2014. Inoltre, la pandemia da SARS-CoV-2 ha contribuito ad accrescere tale tendenza. Secondo

[1] «The *dblp computer science bibliography* provides open bibliographic information on major computer science journals and proceedings. Originally created at the University of Trier in 1993, dblp is now operated and further developed by Schloss Dagstuhl», https://dblp.uni-trier.de/statistics/recordsindblp.html . Consultato il 27/02/2023

[2] PubMed Central® (PMC) is a free full-text archive of biomedical and life sciences journal literature at the U.S. National Institutes of Health's National Library of Medicine (NIH/NLM). In keeping with NLM's legislative mandate to collect and preserve the biomedical literature, PMC serves as a digital counterpart to NLM's extensive print journal collection, https://www.ncbi.nlm.nih.gov/pmc/about/intro/ . Consultato il 27/02/2023

Jeffrey Brainard, sulla sola rivista *Science* sono stati pubblicati circa 23.000 articoli da gennaio a maggio 2020 sul COVID-19[3]. Garantire l'adozione di strategie volte ad assicurare la conservazione digitale e l'accesso ai prodotti della ricerca è essenziale per migliorare la disseminazione della conoscenza scientifica con ricadute positive anche in termini economici. Le diverse misure adottate nel contesto internazionale costituiscono un passaggio importante che lascia spazio ancora a numerose riflessioni anche alla luce della rapida evoluzione tecnologica. Per esempio, un primo elemento critico è costituito dall'ambiguità che talvolta è possibile riscontrare nella definizione dei prodotti della ricerca in rapporto ai contesti applicativi. Pertanto, il primo passo necessario è indagare e tentare di elaborare la definizione di prodotti della ricerca al fine di chiarirne sotto il profilo semantico la reale portata. In generale, è possibile affermare che i prodotti della ricerca sono opere frutto dell'ingegno di ricercatrici e ricercatori generati nel corso del processo di scoperta scientifica. In tal senso, i prodotti della ricerca non rimandano solo alle pubblicazioni scientifiche, ma anche a tutti quegli oggetti informativi che si producono durante la gestione di un progetto di ricerca (il progetto stesso e i suoi allegati, i dati relativi a esperimenti e test, le basi di dati, i software, i prototipi, i brevetti, la manualistica e i documenti tecnici, i report e la corrispondenza elettronica).

Per ciò che attiene alle pubblicazioni, la «crisi dei prezzi» provocata da importanti tagli alle risorse bibliotecarie e la graduale monopolizzazione delle pubblicazioni scientifiche da parte dei grandi editori hanno comportato una riduzione nell'acquisto di risorse scientifiche da parte delle biblioteche limitando disseminazione della conoscenza. Basti pensare che dal 1986 al 2015 l'incremento dei prezzi delle riviste è stato pari al 521% mentre il guadagno netto annuale di *Elsevier* nel 2020 è stato del 38% del PIL globale, contro il 25% di Google[4]. In tale scenario, la comunità internazionale dei ricercatori matura l'esigenza di incentrare la comunicazione internazionale dei ricercatori matura l'esigenza di utilizzare depositi di accesso aperto. Tra le prime esperienze note figura quella realizzata da Paul Ginsparg nel 1991 sull'apertura del repository del Los Alamos National Library contenete i pre-print elettronici di ricerche fisiche e matematiche. Il repository è ancora consultabile[5] e oggi ha esteso l'accesso ai contributi scientifici anche di altri ambiti scientifici inerenti le scienze tecniche. Un'ulteriore esperienza riguarda l'implementazione nel 1999 del protocollo per l'interoperabilità tra repository da parte dei ricercatori e bibliotecari di Santa Fe (USA) al fine di recuperare in modo efficiente e da remoto i contributi scientifici full text battezzato Open Archive Initiative (OAI)[6]. L'evento che segna la svolta nel mondo della comunicazione scientifica è la conferenza *Budapest Open Access*

[3] https://www.sciencemag.org/news/2020/05/scientists-are-drowning-covid-19-papers-can-new-tools-keep-them-afloat . Consultato il 27/02/2023

[4] Elena Giglia, "*Open Science dalla A alla Z – 1- Comunicazione Scientifica Oggi*", link: https://zenodo.org/record/3907297#.X4601UIzZGM . Consultato il 27/02/2023

[5] arXiv, link: https://arxiv.org/ . Consultato il 27/02/2023

[6] OpenArchives, link: http://openarchives.org . Consultato il 27/02/2023

Initiative (BOAI) nel 2002 durante la quale è formalmente adoperato per la prima volta all'interno del manifesto il termine *Open Access*[7]In particolare, il manifesto definisce due strategie alternative note come *green-road* e *gold-road*[8]. Alla BOAI si sono succeduti altri eventi a sostegno dell'Open Access, tra cui la *Bethesda Statement* e la *Berlin Declaration* la cui tripartizione è nota con l'acronimo BBB Definition. Nel 2004 anche l'Italia formalizza la condivisione delle politiche dell'Open Access attraverso la stesura della "*Dichiarazione di Messina. Documento italiano a sostegno della dichiarazione di Berlino sull'accesso aperto alla letteratura accademica"* a cui aderiscono 71 università italiane su 82 fino al 2010[9].

L'importanza che il tema dell'accesso aperto riveste nel tempo anche in termini politici determina la presa di posizione anche dell'Unione Europea. In particolare, la posizione dell'UE rispetto alle politiche dell'accesso aperto muta nel corso del tempo poiché prima tende a disincentivare la disseminazione della conoscenza scientifica per poi adottare soluzioni opposte. Una tra le prime reazioni dell'Unione Europea al tema della scienza aperta è formalizzata nella direttiva n. 98/2003 emanata dal Parlamento europeo «relativa al riutilizzo dell'informazione del settore pubblico» che, secondo quanto stabilito dall'art. 1 comma 1, «detta un complesso minimo di norme in materia di riutilizzo e di strumenti pratici per agevolare il riutilizzo dei documenti esistenti in possesso degli enti pubblici degli Stati membri». Tuttavia, l'art. 1 comma 2 let. e) indica che la direttiva non si applica «e) ai documenti in possesso di istituti d'istruzione e di ricerca quali scuole, università, archivi, biblioteche ed enti di ricerca, comprese, ove opportuno, organizzazioni preposte al trasferimento dei risultati della ricerca»[10]. Successivamente, la direttiva n. 37/2013 introduce una modifica alla precedente direttiva n. 98/2003 sostituendo la let. e) con la seguente dicitura «ai documenti in possesso di istituti di istruzione e di ricerca, comprese organizzazioni preposte al trasferimento dei risultati della ricerca, scuole e università, escluse le biblioteche universitarie». In tal senso, le modifiche hanno introdotto l'obbligo di consentire il riutilizzo dei dati pubblici ampliando l'ambito di applicazione della direttiva anche alle biblioteche universitarie.[11] Infine, alla direttiva n. 37/2013 hanno fatto seguito la direttiva n. 111/2018 - ad integrazione della precedente in cui la Commissione Europea intende analizzare le questioni rimaste irrisolte tra cui quelle volte al «rafforzamento dell'economia dei dati dell'UE grazie a iniziative volte ad aumentare la quantità di dati del settore pubblico messi a disposizione per il riutilizzo, a garantire una concorrenza leale e un facile accesso ai mercati basati sull'informazione del settore pubblico e a

[7] CASSELLA, Maria, *Open Access e comunicazione scientifica*, Editrice Bibliografica, Milano, 2012

[8] GUERRINI, Mauro, *Gli Archivi Istituzionali*, Editrice Bibliografica, Milano, 2010

[9] http://www.sssup.it/UploadDocs/7109_Dichiarazione_di_Messina.pdf. Consultato il 27/02/2023

[10] Direttiva 2003/98/CE del Parlamento europeo e del Consiglio, del 17 novembre 2003, relativa al riutilizzo dell'informazione del settore pubblico. *GU L 345 del 31.12.2003, pag. 90–96*

[11] Ulteriori modifiche hanno riguardato la definizione di una regola standard di tariffazione limitata ai costi marginali per la riproduzione, la messa a disposizione e la divulgazione delle informazioni, e hanno obbligato gli enti pubblici ad assicurare maggiore trasparenza per quanto riguarda le regole di tariffazione e le condizioni da essi applicate

promuovere l'innovazione a livello transfrontaliero basata sui dati»[12] - e la direttiva n. 1024/2019 relativa all'apertura dei dati e al riutilizzo dell'informazione del settore pubblico – in cui si definiscono i termini commerciali del riutilizzo dei dati del settore pubblico i cui diritti di proprietà intellettuale sono detenuti dalle biblioteche universitarie.

A seguito della lettera del 28 aprile 2005 inviata da sei capi di Stato e di governo alla Commissione con la quale si chiedeva di adottare misure necessarie per migliorare l'accesso al patrimonio culturale e scientifico europeo, la stessa emana il 14 febbraio 2007 la «Comunicazione della commissione al parlamento europeo, al consiglio e al comitato economico e sociale europeo sull'informazione scientifica nell'era digitale: accesso, diffusione e conservazione».

In particolare, la comunicazione intende avviare un processo strategico incentrato su due direttrici relative a incentivare a) l'accesso e la diffusione dell'informazione scientifica e migliorare b) le strategie per la conservazione dell'informazione scientifica nell'Unione europea anche alla luce di un'ulteriore importante iniziativa europea avviata per le medesime ragioni contenute nella lettera alla Commissione denominata "Digital Library i2010"[13].

La comunicazione inerente le "Digital Library" intende indicare la strategia finalizzata alla costruzione di una biblioteca europea virtuale attraverso la digitalizzazione, l'accessibilità e la conservazione digitale del patrimonio culturale e scientifico dell'Europa. Una delle esperienze realizzate in tale contesto è l'implementazione del modello di riferimento per le Digital Library DELOS da parte della rete di eccellenza coadiuvata dall'ISTI-CNR di Pisa[14], DELOS costituisce il modello di riferimento per la definizione del framework di politiche, metodologie e tecnologie relativi al dominio applicativo per la conservazione e l'accesso ai contributi della ricerca.

Parallelamente alle azioni legislative, l'Unione Europea ha investito cospicui finanziamenti per la realizzazione di progetti finalizzati all'applicazione delle politiche di apertura dei prodotti della ricerca incentrate sulla strategie dell'Open Science. Anche l'UNESCO definisce la scienza aperta «as an inclusive construct that combines various movements and practices aiming to make multilingual scientific knowledge openly available, accessible and reusable for everyone, to increase scientific collaborations and sharing of information for the benefits of science and society, and

[12]Commissione Europea, "Proposta di Direttiva del parlamento europeo e del consiglio relativa al riutilizzo dell'informazione del settore pubblico*,* Bruxelles, April 25, 2018;

[13]Parlamento europeo, "Comunicazione della Commissione del 30 settembre 2005 al Consiglio, al Comitato economico e sociale europeo e al Comitato delle regioni – i2010: biblioteche digitali", GU C 49 del 28.2.2008], link: https://eur-lex.europa.eu/EN/legal-content/summary/i2010-digital-libraries.html. Consultato il 27/02/2023

[14] «The digital library universe is a complex framework». A Network of Excellence on Digital Libraries. Instrument: Network of Excellence. Thematic Priority: IST-2002-2.3.1.12. Technology-enhanced Learning and Access to Cultural Heritage. "*The DELOS Digital Library Reference Model. Foundations for Digital Libraries*", Version 0.96 November 2007 (http://delosw.isti.cnr.it/files/pdf/ReferenceModel/DELOS_DLReferenceModel_096.pdf. Consultato il 27/02/2023

to open the processes of scientific knowledge creation, evaluation and communication to societal actors beyond the traditional scientific community»[15]. L'Open Science può essere inteso come un insieme di concetti finalizzati alla diffusione della conoscenza scientifica. Tale finalità prevede anche strategie di conservazione digitale e recupero dei prodotti scientifici. Una prima risposta all'attuazione delle politiche Open Science risale al programma Horizon 2020[16] che obbliga i ricercatori che usufruiscono dei finanziamenti europei a pubblicare i risultati delle ricerche secondo le modalità dell'Open Access. Successivamente la Commissione europea avvia nel 2004 la consultazione pubblica denominata "*Science 2.0: Science in Transition*". Obiettivo della consultazione è raccogliere le visioni delle parti interessate dal contesto scientifico al fine di comprendere pienamente l'impatto potenziale della Scienza 2.0. e definire il quadro d'azione per la trasformazione digitale inerente la pratica di diffusione della conoscenza scientifica[17]. In particolare, i risultati della consultazione pubblica mostrano che gli stakeholder preferiscono utilizzare per la prima volta il termine "Open Science" invece di "Science 2.0". La scelta della terminologia è significativa poiché mira a valorizzare il concetto di fruibilità della scienza e di libero accesso alla stessa.

In tal senso, il documento sottolinea che l'adozione dei principi del movimento Open Science sono «l'evoluzione in corso nel modus operandi di fare ricerca e organizzare la scienza»[18]. Per supportare gli stakeholders nell'adozione dei principi dell'Open Science, la Commissione europea stanzia nel 2014 circa 2 milioni di euro per la realizzazione del progetto FOSTER (facilitate open science training for european research)[19] grazie al quale è stato possibile modellare la tassonomia dell'Open Science di seguito illustrata.

[15] UNESCO, "UNESCO Recommendation on Open Science", CC BY-SA 3.0 IGO, 2021, link: https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en . Consultato il 27/02/2023

[16] Unione Europea, "Horizon 2020", link: https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-2020_en . Consultato il 27/02/2023

[17] Commissione Europea, "Validation of the results of the public consultation on Science 2.0: Science in Transition", link: https://www.espci.psl.eu/sites/www.espci.psl.eu/IMG/pdf/science_2_0_final_report.pdf . Consultato il 27/02/2023

[18] Ibidem

[19] Commissione Europea, "Cordis", link: https://cordis.europa.eu/project/id/612425 . Consultato il 27/02/2023

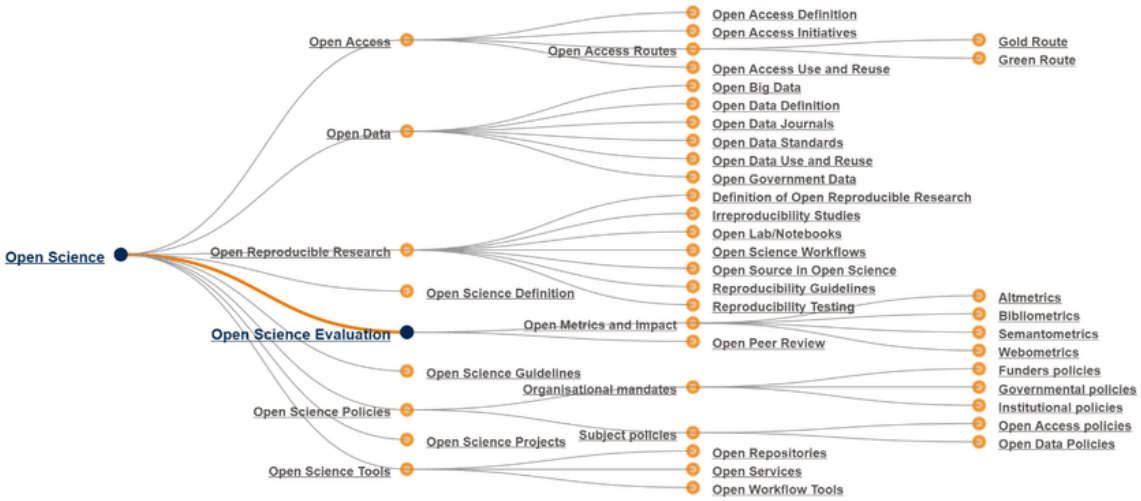

Figura 2. La tassonomia dell'Open Science proposta nel progetto FOSTER (https://cordis.europa.eu/project/id/612425)

In seguito alla consultazione pubblica del 2014, la Commissione Europea identifica cinque linee di potenziali azioni politiche per sostenere lo sviluppo della scienza aperta in Europa. I potenziali interventi si basano sull'aspettativa di rendere la scienza aperta, più credibile, affidabile, efficiente e più reattiva alle sfide della società. In particolare, le cinque linee di potenziali azioni politiche sono:

- Promuovere e creare incentivi per la scienza aperta, promuovendo le migliori pratiche e aumentando il contributo dei produttori di conoscenza in un ambiente più aperto (scienza dei cittadini). Quest'area si occupa anche di garantire la qualità, l'impatto e l'integrità della ricerca della scienza (aperta);

- Rimuovere le barriere all'Open Science: ciò implica, tra le altre questioni, una revisione delle carriere dei ricercatori in modo da creare incentivi e ricompense per l'impegno nell'Open Science;

- Integrare e promuovere ulteriormente le politiche di accesso aperto per quanto riguarda i dati di ricerca e la pubblicazione della ricerca;

- Sviluppo di infrastrutture di ricerca per la scienza aperta, per migliorare l'hosting, l'accesso e la governance dei dati, con lo sviluppo di un quadro comune per i dati di ricerca e la creazione di un cloud europeo per la scienza aperta, un'importante iniziativa per costruire la necessaria infrastruttura di scienza aperta in Europa;

- Integrare la scienza aperta nella società come motore socioeconomico, per cui la scienza aperta diventa determinante nel rendere la scienza più reattiva alle

aspettative sociali ed economiche, in particolare affrontando le principali sfide affrontate dalla società[20].

Ad integrazione delle linee politiche espresse, un'ulteriore misura volta a sviluppare delle proposte finalizzate alla realizzazione delle politiche della scienza aperta nei confronti della Commissione Europea a partire dal 2006 è la Open Science Policy Platform (OSPP, anche EUOSPP)[21]. In tal senso, la piattaforma politica sulla scienza aperta composta da stakeholders di 25 paesi membri fornisce linee di indirizzo alla Commissione per l'attuazione di questioni trasversali riguardanti la scienza aperta in linea con le cinque grandi linee di azione politica. Tra i diversi temi affrontati dai gruppi di esperti figurano l'Open Education and Skills (OES), Citizen Science (CS), la definizione di modelli di business in evoluzione per l'editoria, premi per i ricercatori, altmetrics, integrità della ricerca, e, infine, lo European Open Science Cloud (EOSC) e i l'implementazione di dati FAIR. In particolare, gli ultimi due temi sono intrinsecamente legati poiché l'implementazione di un cloud europeo per la condivisione dei risultati scientifici incentrata sulle politiche di apertura richiede l'adozione di soluzioni che garantiscano i requisiti FAIR dei dati:

- Findable;
- Accessible;
- Interoperable;
- Reusable[22].

Per esempio, GOFAIR segue una strategia di implementazione aperta dal basso verso l'alto per la governance tecnica e i finanziamenti necessari per stabilire la prima fase dell'European Open Science Cloud (EOSC) come parte di un più ampio Internet globale di dati FAIR. Un ruolo cruciale dei dati FAIR della ricerca è ricoperto dai metadati che permettono la descrizione anche ai fini della conservazione delle risorse scientifiche[23]. L'incremento di tali risorse evidenzia ulteriormente l'esigenza di prevedere l'adozione di schemi di metadati che soddisfino i requisiti di accessibilità e recupero nel tempo. Gestire tali metadati è cruciale e, seppur l'implementazione di algoritmi di intelligenza artificiale ai fini della estrazione automatica costituisca una prima risposta alla complessa questione della gestione dei metadati, diviene necessario riflettere su soluzioni alternative e maggiormente innovative che garantiscano l'estrazione anche in contesti distribuiti come il Cloud.

[20] Commissione Europea, "*Open Innovation Open Science Open to the World*. *A vision For Europe*", 2016 P. 45

[21] OpenScience.eu, "Open Innovation Open Science Open to the World. A vision For Europe", link: https://openscience.eu/open-science-policy-platform-final-report. Consultato il 27/02/2023;

[22] WILKINSON, Mark D., DUMONTIER, Michel, AALBERSBERG, IJsbrand, et alii "*The FAIR Guiding Principles for scientific data management and stewardship*", Scientific data, Nature, 3:160018. DOI: 10.1038/sdata.2016.18

[23] Per maggiori dettagli in merito al ruolo dei metadati nella realizzazione dei dati FAIR si veda il par. 1.2. del presente lavoro di tesi

Sulla scia delle iniziative europee, la maggior parte degli stati membri ha adottato misure volte a realizzare nel contesto nazionale le politiche dell'Open science. Una delle misure concerne la redazione del piano nazionale per la scienza aperta adottato tra gli altri da Francia[24], Paesi Bassi[25], Ucraina[26] e, di recente, anche l'Italia[27]. Quest'ultimo è stato redatto da un gruppo di lavoro composto da esperti con competenze eterogenee (fisici e bibliotecari) e illustra il programma nazionale per la ricerca aperta da realizzare entro il quinquennio 2021-2027. In particolare, il piano si sviluppa su cinque assi di intervento per la realizzazione della scienza aperta in Italia suddivisi in:

- *pubblicazioni scientifiche*;
- *dati della ricerca scientifica;*
- *valutazione della ricerca*;
- *scienza aperta* e *apertura dei dati della ricerca su SARS-COV-2* e *COVID-19.*

Per ciascun asse è stato definito il relativo *piano di intervento*. Tra gli obiettivi del secondo asse figura la promozione degli investimenti necessari per la produzione di nuovi dati *FAIR-by-design* "con la generazione automatica, ove possibile, dei metadati e della appropriata informazione contestuale che ne faciliti la ricerca e il riuso"[28], sebbene il ruolo dei metadati possa essere esteso anche a tutti i piani di intervento di ciascun asse. In tal senso, l'adozione del piano nazionale mira a favorire l'accesso e lo scambio informativo tra i ricercatori e gli istituti di ricerca al fine di accelerare il processo conoscitivo incentivato dall'accesso gratuito ai prodotti della ricerca, allargare gli orizzonti di ciascun ambito della ricerca favorendo la interdisciplinarità, verificare i risultati ottenuti per ottimizzare la qualità della ricerca, economicizzare le risorse grazie al riuso e allargare la partecipazione diretta alla scienza al territorio e ai cittadini. Tuttavia, la genericità del piano lascia ancora ampio margine di riflessione anche in termini di accesso e conservazione del materiale scientifico. Ad esempio, la scelta di metodologie di conservazione digitale che prevedano strategie di gestione delle risorse durante l'intero ciclo di vita è fondamentale per garantire la leggibilità nel tempo. Un

[24] Ministère de L'Enseignement supérieur et de la Recherche, "Second National Plan for Open Science", link: https://www.ouvrirlascience.fr/second-national-plan-for-open-science-2021-2024/. Consultato il 27/02/2023

[25] Publications Office of the European Union, "The Netherlands' plan on open science. Open science monitor case study", link: https://op.europa.eu/en/publication-detail/-/publication/20d4026e-4478-11e9-a8ed-01aa75ed71a1/language-en. Consultato il 27/02/2023;

[26] Ministry of Education and Science of Ukraine, "Ukraine has joined the EU countries that have an approved plan for implementing the Open Science principles", link: https://www.kmu.gov.ua/en/news/ukraina-pryiednalas-do-krain-ies-shcho-maiut-zatverdzhenyi-plan-realizatsii-pryntsypiv-vidkrytoi-nauky. Consultato il 27/02/2023;

[27] Ministero dell'Università e della Ricerca, "Piano Nazionale per la Scienza Aperta", link: https://www.mur.gov.it/sites/default/files/2022-06/Piano_Nazionale_per_la_Scienza_Aperta.pdf. Consultato il 27/02/2023;

[28] Ibidem

ruolo cruciale è giocato anche dai metadati definiti sulla base di standard internazionali e rappresentati mediante formati aperti che assicurino il recupero, la leggibilità degli oggetti e l'interoperabilità tra repository.

1.2. Metadati

Il volume e la velocità con cui sono prodotte le risorse in formato digitale anche nel contesto archivistico aumentano quotidianamente in modo esponenziale. La ragione principale risiede nella diffusione capillare delle tecnologie dell'informazione e della comunicazione (*Information and Communication Technologies - ICT*) a qualunque livello in qualsiasi organizzazione. Tale diffusione da un lato determina l'incremento della quantità e dell'eterogeneità delle informazioni disponibili, dall'altro richiede l'adozione di soluzioni adeguate ai fini della loro gestione durante l'intero ciclo di vita per garantire la conservazione, il recupero e la leggibilità. In tal senso, un ruolo centrale è giocato dai metadati. In accordo, con Richard Gartner «etymology is always a good place to start defining a concept and metadata is no exception. The prefix meta - comes from ancient Greek and is usually translated into English by the preposition about. It is often used to express an idea that is in some way self-reflexive [...]. So, unsurprisingly, metadata is usually defined as data about data»[29]. In tal senso, Stefano Vitali definisce un insieme di metadati come una risorsa a corredo delle entità digitali «nel quale specchiarsi, che parli per loro, che ne racconti in qualche modo la storia»[30]. In generale, i metadati forniscono un complesso descrittivo di informazioni strutturate afferenti a qualunque risorsa. Tale complesso si estrinseca in una serie di elementi in rapporto alla tipologia di risorsa trattata, al contenuto informativo, al contesto di produzione e all'obiettivo prefissato. Per esempio, Murtha Baca evidenzia che i metadati associati ai documenti informatici «refer to all the contextual, processing, preservation, and use information needed to identify and document the scope, authenticity, and integrity of an active or archival record in an electronic record-keeping or archival preservation system»[31]. Secondo l'ISO 15489:2016 i metadati associati a un documento sono «data that describing context, content and structure of records and their management through time»[32] in cui il contesto si riferisce ai soggetti, agli eventi, al tempo e allo spazio di creazione del documento, il contenuto si riferisce al messaggio trasmesso attraverso il documento e, infine, la struttura rimanda alla correlazione associativa tra gli elementi che compongono il documento o tra i documenti medesimi. L'identificazione dei metadati per la gestione dei documenti informatici è una questione in continua evoluzione che trova il costante supporto degli organismi internazionali di standardizzazione nell'elaborazione di schemi anche ai fini dell'interoperabilità tra i dati (si pensi per esempio all'ISO 23081-1:2006 Information and documentation - Records management - Metadata for records, Part. 1: Principles al

[29] GARTNER, Richard, "Metadata – Shaping Knowledge from Antiquity to the Semantic Web", in Springer, 2016. DOI:10.1007/978-3-319-40893-4;
[30] VITALI, Stefano, *Passato digitale*, Bruno Mondadori, Milano, 2004. p. 189;
[31] BACA, Murtha, "Introduction to Metadata. Second Edition", in The Getty Research Institute, (CA 90049-1682), 2008. p.7-19;
[32] ISO 15489-1:2016 Information and Documentation – Records Management – Part:1 Concepts and principle

quale segue la pubblicazione nel 2007 della Part.2: Conceptual and implementation issues[33] per identificare i requisiti essenziali nella definizione di un adeguato set di metadati che documenti anche ai fini della conservazione le informazioni relative a ogni fase della vita di un documento elettronico). Tuttavia, tali prerequisiti non sempre sono garantiti. Per esempio, l'aumento dei prodotti della ricerca[34] ha indotto gli istituti di ricerca all'urgente definizione di appositi set di metadati per offrire ai ricercatori di diverse discipline, l'opportunità di accedere e riutilizzare tale patrimonio informativo, migliorare il processo di knowledge discovery, favorire l'analisi interdisciplinare dei fenomeni oggetto d'indagine, ottimizzare gli investimenti pubblici e aprire al territorio le scoperte scientifiche anche in ottica collaborativa tra le realtà imprenditoriali e i centri di ricerca, ma i set minimi proposti non risultano utili a garantire il controllo del valore informativo e la possibilità di accesso alle risorse nel tempo. In tal senso, prevedere la definizione di un adeguato set di metadati nel contesto accademico permetterebbe la gestione dei prodotti della ricerca generati dai ricercatori nel corso dell'attività di ricerca.

Per supportare le università e i centri di ricerca nella definizione di metadati che valorizzino il contenuto informativo e migliorino l'interoperabilità tra repository anche al fine di supportare le pratiche dell'Open Science, nel 2014 sono stati elaborati i principi FAIR[35].

In particolare, tali principi prevedono che per essere:

- Rintracciabili (**F**indable), i dati della ricerca devono:
    - essere ben descritti dai metadati;
    - prevedere l'uso di metadati per l'identificazione univoca e persistente;
    - prevedere la descrizione del sistema di identificazione dei dati adoperato;
    - prevedere la registrazione o indicizzazione dei metadati come risorsa ricercabile.

- Accessibili (**A**ccessible) i dati della ricerca devono:
    - essere recuperabili attraverso i metadati del sistema di identificazione utilizzando un protocollo di comunicazione standardizzato;
    - il protocollo di comunicazione deve essere aperto, gratuito e universalmente implementabile;
    - il protocollo deve prevedere una procedura di autenticazione e autorizzazione, ove necessario;

[33] ISO 23081-1/2:2006-2007 Information and documentation - Records management - Metadata for records, Part. 1: Principles - Part. 2: Conceptual and implementation issues

[34] Per maggiori dettagli in merito si veda il paragrafo 1.1 del presente documento

[35] FORCE11, "The Fair Data Principles", link: https://www.force11.org/group/fairgroup/fairprinciples. Consultato il 27/02/2023

- prevedere l'accessibilità ai metadati anche quando i dati non sono più disponibili.

- Interoperabili (**I**nteroperable) i dati della ricerca devono:
  - Utilizzare un linguaggio descrittivo formale, raggiungibile, condiviso e ampiamente applicabile per la rappresentazione della conoscenza dei metadati;
  - prevedere l'uso di metadati incentrati su un vocabolario che segue i principi FAIR;
  - prevedere l'uso di metadati contenenti riferimenti identificabili ad altri metadati.

- Riutilizzabili (**R**e-usable)i dati della ricerca devono:
  - prevedere l'uso di metadati ricchi di attributi, pertinenti e accurati;
  - prevedere l'uso di metadati pubblicati che indichino le licenze di utilizzo dei dati in modo chiaro e accessibile;
  - prevedere l'uso di metadati collegati al contesto di produzione;
  - prevedere l'uso di metadati che soddisfano gli standard della comunità in ogni campo[36].

La definizione di un adeguato set di metadati dovrebbe inoltre permettere l'aggregazione dei prodotti della ricerca ai documenti amministrativi informatici generati in funzione dell'erogazione del servizio di ricerca. Dal punto di vista amministrativo, la recente entrata in vigore delle Linee guida contenenti regole tecniche attuative del D. Lgs 7 marzo 2005 n. 82, recante il *Codice dell'Amministrazione Digitale,* emanate da Agenzia per l'Italia Digitale, secondo quanto disposto dall'art. 71 dello stesso Codice[37], stabiliscono al paragrafo 2.1.1. «al momento della formazione del documento informatico immodificabile, devono essere generati e associati permanentemente ad esso i relativi metadati»[38] indicati all'interno dell'allegato 5 denominato "Metadati". In particolare, tale allegato distingue tra metadati da associare al documento informatico, al documento amministrativo informatico e all'aggregazione

[36] Ibidem

[37] Art. 71 del CAD stabilisce che: "L'AgID, previa consultazione pubblica da svolgersi entro il termine di trenta giorni, sentiti le amministrazioni competenti e il Garante per la protezione dei dati personali nelle materie di competenza, nonché acquisito il parere della Conferenza unificata, adotta Linee guida contenenti le regole tecniche e di indirizzo per l'attuazione del presente Codice. Le Linee guida divengono efficaci dopo la loro pubblicazione nell'apposita area del sito Internet istituzionale dell'AgID e di essa ne è data notizia nella Gazzetta Ufficiale della Repubblica italiana. Le Linee guida sono aggiornate o modificate con la procedura di cui al primo periodo", https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legislativo:2005-03-07;82 . Consultato il 27/02/2023

[38] Paragrafo 2.1.1. "*Formazione del documento informatico*" in "*Linee Guida sulla formazione, gestione e conservazione dei documenti informatici*" emanate da Agenzia per l'Italia Digitale, https://www.agid.gov.it/sites/default/files/repository_files/linee_guida_sul_documento_informatico.pdf . Consultato il 27/02/2023

documentale informatica[39]. Rispetto all'allegato 5 delle precedenti regole tecniche[40], i metadati delle Linee Guida sono stati ampliati da 5 a 18 elementi. Per ogni metadato viene richiesto di indicare:

- **informazione**: il nome;
- **sottocampi**: l'eventuale sottostruttura del metadato complesso;
- **valori ammessi**: valori accettati all'interno del campo;
- **tipo dato**: numerici o alfanumerici;
- **obbligatorietà**: l'indicazione di obbligatorietà, eventualmente condizionata;
- **nuova definizione**: metadati nuovi o ridefiniti rispetto all'allegato alla normativa precedente;
- **Definizione**: indicazioni sulla modalità di utilizzo del metadato.

Convenzionalmente la letteratura sui metadati spesso rimanda ad una classificazione degli che è organizzata come di seguito:

- *descrittivi*;
- a*mministrativi/di processo;*

La prima classe è quella dei metadati descrittivi e fornisce le informazioni necessarie all'utente per individuare e recuperare una risorsa. In particolare, tra i metadati descrittivi figurano quelli utili all'identificazione univoca e persistente che rappresenta uno degli aspetti fondamentali del recupero in ambiente digitale.

La seconda classe di metadati utile a rappresentare la risorsa durante il suo ciclo di vita è quello dei metadati amministrativi. Tale classe si presenta come articolata e complessa in quanto definisce anche le informazioni necessarie per permettere la gestione, la conservazione e l'accesso nel tempo agli oggetti digitali. Talvolta i metadati amministrativi sono suddivisi in metadati tecnici, di conservazione e per la gestione dei diritti.

[39] L'aggregazione documentale informatica è definita dall'Allegato 1 "Glossario dei termini e degli acronimi" come "Insieme di documenti informatici o insieme di fascicoli informatici riuniti per caratteristiche omogenee, in relazione alla natura e alla forma dei documenti o in relazione all'oggetto e alla materia o in relazione alle funzioni dell'ente", https://www.agid.gov.it/sites/default/files/repository_files/allegato_1_glossario_dei_termini_e_degli_acronimi.pdf . Consultato il 27/02/2023

[40] Il DPCM 3 dicembre 2013 recante "Regole tecniche in materia di sistema di conservazione" abrogato il 1° gennaio 2022

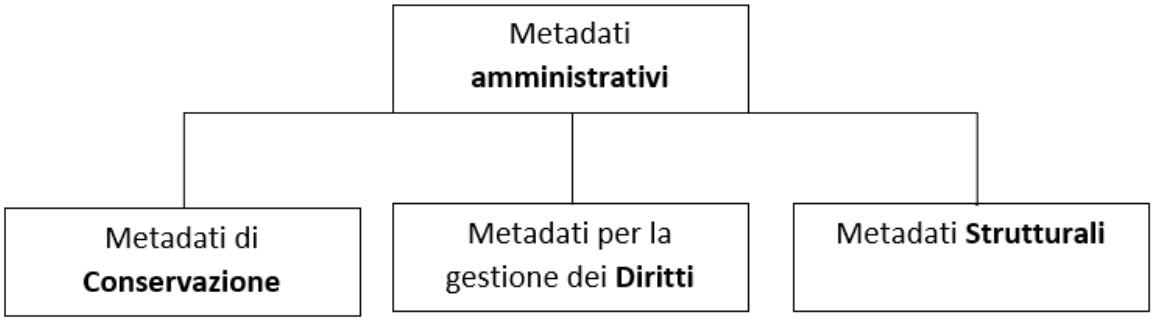

Figura 3. Possibile classificazione dei metadati amministrativi

Uno dei principali rischi legati alla diffusione delle tecnologie dell'informazione e della comunicazione è l'accessibilità delle risorse digitali nel lungo periodo. Si tratta di un rischio strettamente correlato all'obsolescenza tecnologica determinata dall'utilizzo di formati proprietari e dalla dipendenza da architetture hardware e software nel tempo deprecate. Per mitigare tale rischio è necessario prevedere misure adeguate al fine di garantire la riproducibilità nel tempo. Una delle misure fondamentali adoperate ai fini della conservazione digitale è quella di documentare dettagliatamente le informazioni minime necessarie affinché una risorsa sia leggibile anche agli utenti futuri. In tal senso, i metadati svolgono un ruolo fondamentale consentendo la rappresentazione delle informazioni che possono documentare l'oggetto anche in relazione alle eventuali modifiche che lo stesso subisce sul piano tecnologico.

I metadati utili alla gestione dei diritti all'accesso all'oggetto digitale sono anche essi fondamentali rispetto alla importante proprietà della possibilità d'uso che un archivio digitale deve garantire. In particolare, i diritti relativi all'accesso esplicitati fino a livello dei permessi di lettura/scrittura sono definiti in relazione al contesto nel quale il documento si forma, viene acquisito o gestito e conservato.

Nel contesto giuridico europeo il regolamento n. 679/2016 (anche noto con l'acronimo *GDPR - General Data Protection Regulation*) stabilisce le misure relative al trattamento e alla protezione dei dati personali con riguardo alle persone fisiche, nonché alla libera circolazione di tali dati applicate ai Paesi dello spazio giuridico dell'Unione Europea.

Infine, in relazione alla classificazione citata, l'ultima classe di metadati amministrativi è rappresentata dai metadati tecnici o strutturali. Tutte le classi analizzate in precedenza sono indispensabili per garantire l'adeguata gestione del ciclo di vita di una risorsa digitale. Tuttavia, la classe dei metadati strutturali svolge una funzione centrale ai fini della conservazione digitale. In generale, i metadati strutturali rappresentano le informazioni necessarie per ricostruire la struttura di un oggetto digitale. Dunque, essi sono in grado di rendere manifeste le istruzioni sull'organizzazione ordinata del contenuto informativo e sulla rappresentazione formale mostrata all'utente finale.

Lo schema illustrato di seguito rappresenta una tabella riepilogativa delle classi dei metadati appena descritte.

Classe		**Definizione**	**Funzione**
Descrittivi		Supportano l'individuazione, la scoperta e il recupero di una risorsa	Identificazione; descrizione; ricerca; recupero; contestualizzazione geo-temporale.
Amministrativi/di processo	(di) **conservazione**	Forniscono le informazioni necessarie per garantire l'accessibilità e la leggibilità di una risorsa nel tempo.	Accessibilità; Leggibilità.
	(per la gestione dei) **Diritti**	Stabiliscono le limitazioni degli utenti in rapporto ai diritti dettati dalla normativa di riferimento e dall'organizzazione interna di ciascun soggetto produttore	Protezione dei dati personali; Limitazioni alle azioni eseguibili all'interno di un sistema informativo.
	Strutturali	Contengono le informazioni necessarie per collegare per ricostruire la struttura e l'organizzazione di un oggetto digitale complesso	Formazione della risorsa digitale; Organizzazione; Convalida del valore probatorio.

Figura 4. La tabella riassume l'organizzazione delle diverse classi di metadati, la definizione e la relativa funzione.

1.2.1. Standard internazionali di metadati

I metadati, anche alla luce di quanto sinora affermato, consentono di strutturare la conoscenza. Conseguentemente, nel corso del tempo, sono stati elaborati numerosi

standard internazionali per la definizione, la gestione e l'implementazione dei metadati. Il proliferare degli standard in questo specifico ambito è determinato anche dalla stretta relazione che i metadati mostrano con il dominio applicativo. Pertanto, si è ritenuto di costruire modelli diversificati a seconda del contesto di applicazione. Tuttavia, il numero elevato di standard di metadati ha finito per rappresentare un ostacolo alla metadatazione piuttosto che un ausilio. Gli standard ad oggi elaborati sono riconducibili non solo all'ISO (Organismo internazionale di standardizzazione) ma anche ad altre importanti organizzazioni deputate da sempre all'emanazione di regole e metodologie nel campo della gestione e conservazione documentale[41].

Nel presente lavoro di tesi si è scelto di presentare rapidamente gli standard internazionali utili ai temi qui richiamati mentre si è deciso di non trattare altri importanti modelli, tuttavia, non funzionali all'obiettivo del lavoro.

1.2.1.1. Standard internazionali di metadati descrittivi

Tra gli standard di metadati descrittivi sono annoverati molti modelli alcuni dei quali maggiormente noti per via della loro notevole diffusione. È il caso, ad esempio, del Dublin Core la cui applicazione è andata ben oltre i confini della biblioteconomia oppure degli standard di descrizione archivistica (ISAD-G, ISAAR-Cpf) che attraverso le loro strutturale formali in XML (EAD, EAC) costituiscono un riferimento univoco per gli archivi storici.

Tra gli standard internazionali di metadati descrittivi l'EAD (Encoded Archival Description) è uno tra quelli recentemente impiegati nelle esperienze di conservazione degli archivi digitali. La nascita dello standard risale alla creazione del progetto Berkeley Findings Aids Projects (BFAP) guidato dalla Berkeley University of California e presentato il 1993 in occasione della riunione annuale della Society of American Archivists presso New Orleans. Obiettivo iniziale del progetto BFAP è l'elaborazione di uno standard aperto per la codifica dei materiali archivistici incentrato sull'adozione dello Standard Generalised Mark-up Language (SGML) per supportare la ricerca e l'interoperabilità tra i dati. La versione 1.0 dello standard EAD pubblicata il 1998 conduce all'elaborazione di una DTD (Document Type Declaration) anche grazie al coinvolgimento attivo nel progetto di istituzioni che costituiscono un punto di riferimento del contesto archivistico internazionali come la Society of American Archivists e la Library of Congress di Washington. Nel tempo, l'interesse nei confronti dello standard conduce al coinvolgimento di ulteriori istituzioni, come la Research Libraries Group (RLG) che nel 2002 pubblica[42] delle linee guida sulle *best practies* per

[41] O*nline Computer Library Center* (OCLC), https://www.oclc.org/en/home.html . Consultato il 27/02/2023, la *Library of Congress,* https://www.loc.gov/ . Consultato il 27/02/2023. *Research Libraries Group* (RLG), https://www.rlg.org/ . . Consultato il 27/02/2023. *Society of American Archivist,* https://www2.archivists.org/ . Consultato il 27/02/2023 e UNI (*Ente nazionale italiano di unificazione*), https://www.uni.com/index.php . Consultato il 27/02/2023

[42] Online Computer Library Center, "RLG Best practice Guidelines for Encoded Archival Description", link: https://www.oclc.org/content/dam/research/activities/ead/bpg.pdf. Consultato il 27/02/2023

l'adozione dello standard EAD, e alla revisione dello schema fino all'ultima versione 3.0 rilasciata nel 2015 e basata sullo standard XML (eXtensible Mark-up Language)[43]. L'aggiornamento dello standard è attualmente di responsabilità del sottocomitato tecnico per gli standard per la codifica degli archivi (Technical Subcommittee on Encoded Archival Standards (TS-EAS) della Society of American Archivists in collaborazione con la Library of Congress e grazie all'apertura delle sue specifiche tecniche e all'adozione di standard indipendenti da tecnologie hardware e software (come XML) è adottato da diverse istituzioni archivistiche nazionali (tra cui Stati Uniti, Regno Unito, Francia, Australia e Canada) per la descrizione e pubblicazione in formato elettronico di strumenti per la ricerca archivistica in origine prodotti su supporto cartaceo. Inoltre, la struttura EAD incentrata su 146 elementi[44] può essere matchata con quella di ulteriori standard di metadati come Dublin Core, MARC, ISAD, ISAG e DACS migliorando la complessità di informazioni e ottimizzando l'interoperabilità tra le risorse.

Alcune esperienze richiamano l'applicazione dell'EAC, uno schema XML per la codifica e l'interscambio di record di autorità basato sull'International Standard Archival Authority Record for Corporate Body, Persons and Families(CPF). Lo standard è originariamente pubblicato nella sua versione beta il 2004 e nell'ultima versione il 2022 come EAC-CPF 2.0. In particolare, la struttura dello standard è basata su due elementi obbligatori:

- <control> che consiste nell'acquisizione dei dati relativi alla produzione e all'aggiornamento del record d'autorità;
- <cpfDescription>, che include le informazioni sull'autorità descritta, le relazioni con altre autorità e con le relative risorse connesse o, nel caso di autorità complesse strutturate in molteplici identità può essere impiegato in alternativa l'elemento <multipleIdentities>.

L'Encoded Archival Guide (EAG)[45] costituisce un ulteriore standard del quadro internazionale Encoded Archival Standards (EAS) e uno degli standard richiamati anche nelle esperienze sulla conservazione degli archivi digitali. Supervisionato dal Working Group on Standards (WGoS) dell'Archives Portal Europe Foundation (APEF). Lo sviluppo dell'EAG è coordinato in stretta collaborazione con il TS-EAS, responsabile anche degli standard EAD e dell'EAC-CPF sopracitati e l'attuale versione risale al 2012, elaborata nel corso del progetto APEx. In particolare, l'EAG è uno standard XML sviluppato con l'intento di codificare le informazioni relative agli enti archivistici censiti all'interno

[43] The Library of Congress, "<ead>. Encoded Archival Description", link: https://www.loc.gov/ead/. Consultato il 27/02/2023

[44] Wikipedia, "Encoded Archival Description", link: https://en.wikipedia.org/wiki/Encoded_Archival_Description#History. Consultato il 27/02/2023

[45] Archival Portal Europe, "The use of EAG in Archives Portal Europe", link: https://www.archivesportaleurope.net/tools/for-content-providers/standards/eag/, Consultato il 27/02/2023

dell'Archives Portal Europe[46] il quale offre un unico punto di accesso per recuperare e scoprire informazioni sugli archivi di migliaia di istituzioni del patrimonio culturale appartenenti ad oltre 30 paesi europei.
Se gli standard di descrizione archivistica richiamati hanno avuto già un'ampia applicazione negli archivi digitalizzati e pertanto sono stati utilizzati anche all'interno di processi di conservazione di tali complessi documentali, nel caso degli archivi della ricerca è necessario, come ricordato, definire processi di gestione e di conservazione degli oggetti digitali fin dal momento della loro formazione. In tal caso, è evidente che non si ricorrerà a standard di descrizione archivistica ma piuttosto a standard di descrizione che unitamente all'applicazione di standard di metadati di processo consentiranno la gestione, la tutela e l'esibizione dei documenti durante il loro ciclo di vita. Vale la pena dunque richiamare in questa trattazione lo standard Dublin Core utilizzato anche dal quadro normativo italiano per la descrizione dei documenti informatici e dei documenti amministrativi informatici.

Il modello Dublin Core[47] è stato sviluppato dalla OCLC con la finalità di creare strumenti per la rappresentazione e la condivisione delle risorse digitali. La prima versione pubblicata nel 1996 si componeva di 15 elementi (Core). Successivamente, gli elementi di base sono stati arricchiti con un set di qualificatori in grado di migliorare sul piano semantico le performance degli elementi di metadati relativamente alla rappresentazione degli oggetti in un ambiente digitale. Le caratteristiche di flessibilità, semplicità e interoperabilità hanno determinato nel tempo il successo del Dublin Core che nel 2009 diventa standard ISO 15836[48]. Lo sviluppo continuo di set, application profile, vocabulary, rendono lo standard sempre attuale e adeguato anche al rapido mutare di situazioni e domini come solitamente accade in ambiente digitale.

1.2.1.2. Standard internazionali di metadati amministrativi per la conservazione

Uno degli standard internazionali di riferimento per la definizione dei metadati per la conservazione degli oggetti digitali è il PREMIS (*PREservation Metadata Implementation Strategies*) Data Dictionary for Preservation Metadata[49]. Elaborato per la prima volta nel 2005 grazie alla collaborazione tra la Online Computer Library Center (OCLC) e il Research Libraries Group (RLG) il documento è stato sottoposto a diverse revisioni fino all'ultima risalente al 2015. Parallelamente al Data Dictionary, Premis prevede anche uno schema XML per la rappresentazione periodicamente aggiornato dalla Library of Congress mediante la *PREMIS Maintenance Activity*. In particolare, lo

[46] Archival Portal Europe, "The history of Europe - one click away", link: https://www.archivesportaleurope.net/. Consultato il 27/02/2023
[47] Dublin Core Initiative, https://www.dublincore.org/ . Consultato il 27/02/2023
[48] ISO 15836:2009 Information and documentation — The Dublin Core metadata element set
[49] The Library of Congress, "PREMIS. Preservation Metadata Maintenance Activity", link: https://www.loc.gov/standards/premis/. Consultato il 27/02/2023

standard PREMIS è strutturato in quattro unità semantiche reciprocamente correlate. Ogni unità semantica corrisponde nel Data Dictionary a una entità del data model. Le quattro unità semantiche sono:

- **Oggetti (Objects)**: è l'entità che si riferisce all'oggetto (digitale) inteso come un insieme finito di informazioni che costituiscono l'elemento sottoposto a conservazione digitale;
- **Eventi (Events)**: è l'entità che si riferisce ad una o più azioni che riguardano o l'oggetto digitale sottoposto a conservazione o a un agente noto al sistema di conservazione;
- **Agenti (Agents)**: è l'entità che si riferisce ad una persona, organizzazione o software in relazione agli *Eventi* che si realizzano nel corso del ciclo di vita di un *Oggetto* in rapporto ai *Diritti* posseduti;
- **Diritti (Rights Statement)**: è l'entità che si riferisce al permesso o ai permessi relativi a un *Oggetto* e/o ad un *Agente.*

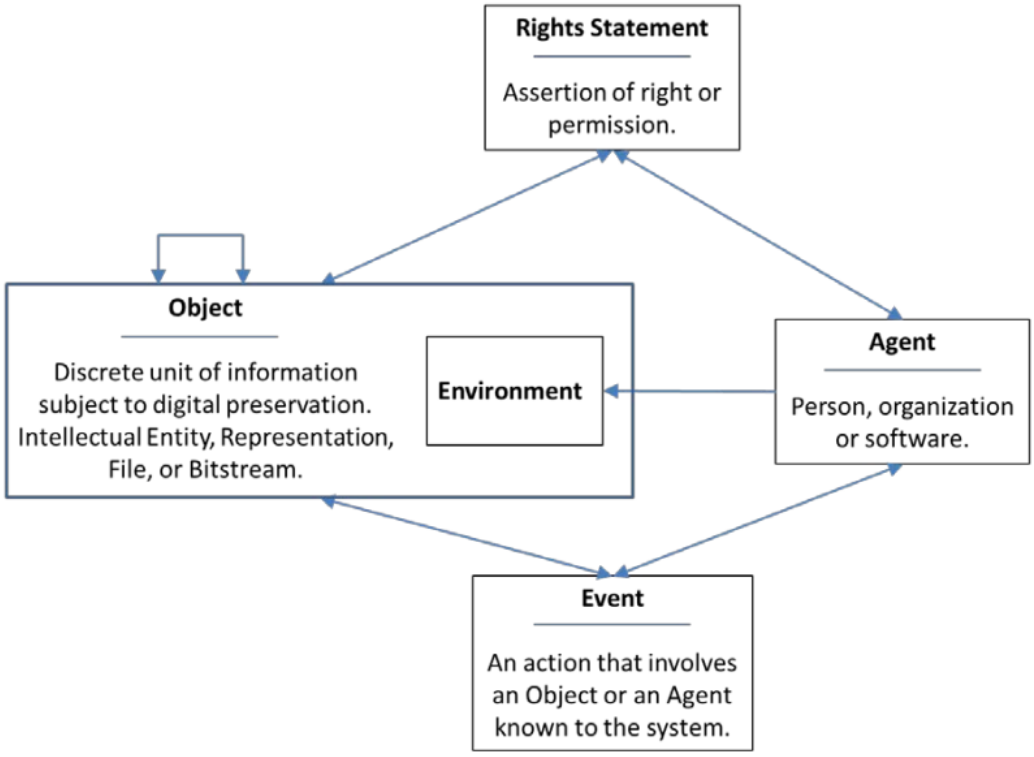

Figura 5. Il modello dei dati PREMIS (Library of Congress Network Development and MARC Standards Office, "Capire PREMIS", link: https://www.loc.gov/standards/premis/Understanding-PREMIS_italian.pdf. Consultato il 27/02/2023;

Lo standard internazionale di metadati PREMIS per la conservazione degli oggetti digitali costituisce uno strumento centrale per gli operatori di istituti archivistici e bibliotecari. Per offrire una panoramica sullo standard di metadati per la conservazione e supportare l'attività di identificazione dei metadati per la conservazione delle risorse digitali

informatiche, il 1° febbraio 2009 la Library of Congress Network Development e il MARC Standards Office pubblicano la guida "*Understanding PREMIS*"[50].

Tale guida indica a titolo esemplificativo le principali attività di conservazione e i relativi metadati tra cui:

- L'identificazione dei metadati (per esempio il c*hecksum*) per assicurare un livello di sicurezza della risorsa conservata tale che si possa dimostrare l'eventuale modifica (intenzionalmente o inavvertitamente) in un determinato arco temporale;
- L'identificazione dei metadati per gestire le informazioni relative alle tipologie di supporti di conservazione delle risorse e le date durante le quali sono state eseguite attività di refresh dei file. In particolare, tali informazioni sono essenziali per assicurare la leggibilità di file tradizionalmente archiviati in passato su supporti magnetici obsoleti (per esempio all'uso allargato dei floppy disk da 8 pollici negli anni '70), ma ancora giuridicamente validi o destinati alla conservazione permanente;
- L'identificazione dei metadati relativi alle strategie adoperate per assicurare la conservazione delle risorse digitali, tra cui le informazioni sulla migrazione da un formato deprecato a un formato aggiornato o le informazioni relativa alla riproduzione dell'ambiente originale hardware e software di creazione per l'emulazione della risorsa;
- L'identificazione dei metadati che documentino la provenienza, le azioni di conservazione e le eventuali evoluzioni necessari per dimostrare l'autenticità della risorsa[51].

Ai fini della conservazione degli archivi digitali, l'utilizzo del modello dei dati PREMIS rappresenta un supporto essenziale per preservare le relazioni intercorse tra gli agenti (identificati dal sistema attraverso l'uso di credenziali di accesso univoche tra cui la password), gli eventi (documentabili grazie alle procedure dei *trigger*) realizzati o non realizzati in rapporto all'oggetto digitale (l'archivio digitale) e ai diritti posseduti (definibile attraverso le *Access-control List*). Tali informazioni costituirebbero l'insieme dei metadati da associare all'archivio digitale e trasferire il complesso di elementi in conservazione. Tale processo di conservazione digitale potrebbe basarsi su un altro standard internazionale definito OAIS (*Reference Model for an Open Archival Information System*) elaborato dal CCSDS (Consultative Committee for Space Data System) e successivamente approvato come standard internazionale ISO 14721:2012 che prevede la conservazione di un oggetto digitale mediante pacchetti informativi (*information packages*) composti dal contenuto informativo (*content information*) che

[50] Library of Congress Network Development and MARC Standards Office, "Capire PREMIS", link: https://www.loc.gov/standards/premis/Understanding-PREMIS_italian.pdf. Consultato il 27/02/2023
[51] Ibidem

include l'oggetto dati (*data object*) che può essere fisico (*physical object*) o digitale (*digital object*) e le informazioni sulla rappresentazione (*representation information*)[52].

Un ulteriore standard di metadati amministrativi è il METS (Encoding And Trasmission Standard)[53]. Elaborato per la prima volta dalla *Society of American Archivist* nel 1998 lo schema è giunto alla sua ultima versione nel 2015 la cui revisione è oggi sotto la responsabilità della Digital Library Federation (DLF). Come evidenziato all'interno del manuale di riferimento dello standard METS redatto e pubblicato dalla DLF «Metadata Encoding and Transmission Standard (METS) is a data encoding and transmission specification, expressed in XML, that provides the means to convey the metadata necessary for both the management of digital objects within a repository and the exchange of such objects between repositories (or between repositories and their users)»[54]. Dal punto di vista strutturale, un documento METS si compone di sette sezioni principali, che possono contenere un'ulteriore varietà di elementi e attributi di seguito indicati:

- **Header**: contiene metadati che descrivono il documento METS, comprese informazioni come creatore e editore;
- **Descriptive Metadata Section:** contiene metadati descrittivi esterni al documento METS (ad esempio, un record MARC in un OPAC o un record MODS mantenuto su un server WWW), metadati descrittivi embedded o entrambi. È possibile includere più istanze di metadati descrittivi esterni e interni nella sezione dei metadati descrittivi;
- **Administrative Metadata Section:** permette di documentare le modalità di creazione e archiviazione dei file, i diritti di proprietà intellettuale, le informazioni relative all'oggetto digitale primario da cui deriva l'oggetto digitale e le informazioni sulla provenienza dei file che compongono l'oggetto (ad esempio, relazioni tra file master/derivati e migrazione/informazioni sulla trasformazione). Come per i metadati descrittivi, i metadati amministrativi possono essere esterni al documento METS o codificati internamente;
- **File Section:** indica l'elenco dei file relativi alle diverse versioni dell'oggetto digitale. I file elements possono essere raggruppati all'interno degli elementi File Group, per suddividere i file in base alla versione dell'oggetto o ad altri criteri come il tipo di file e la dimensione;
- **Structural Map:** Tale sezione delinea la struttura gerarchica dell'oggetto digitale e collega gli elementi di tale struttura ai file di contenuto e ai metadati

[52] Per maggiori dettagli in merito allo standard ISO 14721:2012 OAIS si veda il par. 2.2. del presente lavoro

[53] Library of Congress Network Development and MARC Standards Office, "METS. Metadata Encoding & Trasmission Standard Official Web site" link: https://www.loc.gov/standards/mets/. Consultato il 27/02/2023

[54] Library of Congress Network Development and MARC Standards Office, "Metadata encoding and transmission standard: primer and reference manual", link: https://www.loc.gov/standards/mets/METSPrimer.pdf. Consultato il 27/02/2023;

relativi a ciascun elemento. Questa è l'unica sezione obbligatoria in un documento METS;

- **Structural Links:** fornisce indicazioni in merito a eventuali collegamenti ipertestuali tra i nodi nella gerarchia delineata nella mappa strutturale;
- **Behavior Section:** tale sezione può essere utilizzata per associare operazioni eseguibili al contenuto dell'oggetto codificato secondo le specifiche utilizzate in METS.

Lo standard METS permette di rappresentare i metadati utili all'interscambio tra repository e generalmente viene utilizzato in associazione al Dublin Core e a specifici protocolli per l'ingestion dei pacchetti informativi. Anche i pacchetti di versamento e quelli di distribuzione che caratterizzano il modello OAIS ricorrono allo standard METS per le procedure di versamento e di interoperabilità.
Infine, tra gli standard internazionali per i metadati di processo è importante ricordare anche il modello ISO 23081. Si tratta di uno standard in continuo aggiornamento nelle sue diverse parti che nella articolazione complessa della sua struttura mira a trasferire regole e metodologie per la definizione di modelli e di set di metadati di processo specifici per i diversi domini. Il punto di forza dello standard 23081 è la sua capacità estrema di permettere la rappresentazione di tutte le possibili relazioni interne ed esterne alle diverse entità che il modello prevede. Pertanto, l'applicazione dello standard 23081 consente di rappresentare gli archivi digitali nella loro pienezza di nodi e relazioni che costituiscono la vera ricchezza informativa implicita dei complessi documentali.

1.3. Il quadro normativo di riferimento

1.3.1. Il quadro normativo nazionale

A seguito delle misure promulgate dall'Unione Europea, l'Italia ha elaborato l'*Agenda Digitale Italiana* (ADI) che definisce le strategie di sviluppo, crescita e innovazione abilitate dalle tecnologie digitali. In particolare, la trasformazione digitale della Pubblica Amministrazione è un compito affidato all'Agenzia per l'Italia Digitale (AgID) attraverso la *Strategia per la Crescita Digitale 2014-2020* e il *Piano Triennale per l'informatica nella Pubblica amministrazione*. Parallelamente, il legislatore ha emanato il decreto legislativo 7 marzo 2005 n. 82 recante il *Codice dell'Amministrazione Digitale* (CAD)[55]. L'obiettivo del CAD è racchiudere all'interno di un testo normativo unico il complesso delle disposizioni relative all'informatizzazione di ogni ambito della Pubblica Amministrazione, dall'identità digitale (incluse la firma elettronica, firma elettronica avanzata, firma elettronica qualificata e la firma digitale) alla posta elettronica certificata, dal documento informatico alle copie (informatiche di documenti informatici e analogiche di documenti informatici) e al relativo valore probatorio, dalla formazione, acquisizione, registrazione, protocollazione alla conservazione digitale. Secondo quanto previsto dall'art. 71 comma 1 dello stesso CAD, l'AgID ha emanato le linee guida attuative del Codice sulla formazione, gestione e conservazione dei documenti informatici entrate in vigore a partire dal 1° gennaio 2022. Dal punto di vista strutturale, le Linee Guida si compongono di una base normativa statica e di sei allegati tecnici[56] redatte secondo un approccio olistico al fine di organizzare il complesso di norme, emanate in precedenza separatamente, all'interno di un unico testo normativo. In particolare, il testo si suddivide in quattro capitoli che rispettivamente disciplinano (1) gli strumenti di lettura e le disposizioni comuni, (2) la formazione del documento informatico, del documento amministrativo informatico e il relativo valore probatorio, (3) la gestione documentale e, infine (4) la conservazione.

[55] Parlamento italiano, "D. lgs. recante il Codice dell'Amministrazione Digitale", 7 marzo 2005, n. 82 (in G.U. n.112 del 16-05-2005 - Suppl. Ordinario n. 93)

[56] In particolare, gli allegati si suddividono in:

- Agenzia per l'Italia Digitale, "Allegato 1. Glossario dei Termini e degli Acronimi" in Linee Guida sulla formazione, gestione e conservazione dei documenti informatici, Roma, maggio 2021;
- Agenzia per l'Italia Digitale, "Allegato 2. Formati di file e riversamento" in Linee Guida sulla formazione, gestione e conservazione dei documenti informatici, Roma, 2021;
- Agenzia per l'Italia Digitale, "Allegato 3. Certificazione di processo" in Linee Guida sulla formazione, gestione e conservazione dei documenti informatici, Roma, 2021;
- Agenzia per l'Italia Digitale, "Allegato 4. Standard e specifiche tecniche" in Linee Guida sulla formazione, gestione e conservazione dei documenti informatici, Roma, 2021;
- Agenzia per l'Italia Digitale, "Allegato 5. Metadati" in Linee Guida sulla formazione, gestione e conservazione dei documenti informatici, Roma, 2021;
- Agenzia per l'Italia Digitale, "Allegato 6. Comunicazione tra AOO di Documenti Amministrativi Protocollati" in Linee Guida sulla formazione, gestione e conservazione dei documenti informatici, Roma, 2021.

In particolare, l'entrata in vigore delle Linee Guida ha comportato significative novità rispetto alle precedente regole tecniche[57]. Innanzitutto, la possibilità da parte di AgID di irrogare sanzioni nei confronti dei soggetti indicati nell'art. 2 comma 2 del CAD[58] anche in caso di mancata ottemperanza alle disposizioni delle Linee Guida secondo l'art. 18-bis del CAD. Inoltre, sebbene si tratti di norme di rango non primario, anche i soggetti privati sono sanzionabili secondo quanto previsto dalla normativa primaria in relazione al contesto giuridico di riferimento[59]. Un ulteriore novità concerne i ruoli individuati nel processo di conservazione. Infatti, il precedente DPCM sulla conservazione suddivideva i ruoli in Produttore, Utente e Responsabile della conservazione mentre le nuove Linee Guida specificano che il:

- Titolare dell'oggetto è il soggetto produttore pubblico o privato dei documenti informatici;
- Produttore dei PdV è la figura fisica interna al Titolare dell'oggetto di conservazione che produce il pacchetto di versamento ed è responsabile del trasferimento del suo contenuto nel sistema di conservazione che, nelle pubbliche amministrazioni, si identifica con il responsabile della gestione documentale;
- Responsabile della conservazione è il soggetto interno al Titolare dell'oggetto di conservazione che definisce e attua le politiche complessive del sistema di conservazione e ne governa la gestione con piena responsabilità ed autonomia;
- Utente abilitato è la persona, ente o sistema che interagisce con i servizi di un sistema di gestione informatica dei documenti e/o di un sistema perla

[57] Si fa riferimento alle seguenti disposizioni legislative:

- Presidente del Consiglio dei Ministri, "Regole tecniche per la generazione, apposizione e verifica delle firme elettroniche avanzate, qualificate e digitali", 22 febbraio 2013 (in Gazzetta Ufficiale il 21-5-2013 n. 117);
- Presidente del Consiglio dei Ministri, "Regole tecniche in materia di sistema di conservazione", 3 dicembre 2013 (in Gazzetta Ufficiale il 12-3-2014 n. 59);
- Presidente del Consiglio dei Ministri, "Regole tecniche per il protocollo informatico", 3 dicembre 2013 (in Gazzetta Ufficiale il 12-3-2014 n.59);
- Presidente del Consiglio dei Ministri, "Regole tecniche in materia di formazione, trasmissione, copia, duplicazione, riproduzione e validazione temporale dei documenti informatici nonché di formazione e conservazione dei documenti informatici delle pubbliche amministrazioni", 13 novembre 2014 (in Gazzetta Ufficiale il 12-01-2015 n. 8).

[58] In particolare, le sanzioni sono irrogabili verso:

a) le pubbliche amministrazioni di cui all'articolo 1, comma 2, del decreto legislativo 30 marzo 2001, n. 165, nel rispetto del riparto di competenza di cui all'articolo 117 della Costituzione, ivi comprese le autorità di sistema portuale, nonché alle autorità amministrative indipendenti di garanzia, vigilanza e regolazione;
b) i gestori di servizi pubblici, ivi comprese le società quotate, in relazione ai servizi di pubblico interesse;
c) le società a controllo pubblico, come definite nel decreto legislativo 19 agosto 2016, n. 175, escluse le società quotate di cui all'articolo 2, comma 1, lettera p), del medesimo decreto che non rientrino nella categoria di cui alla lettera b).

[59] Agenzia per l'Italia Digitale, "FAQ", link: https://www.agid.gov.it/it/domande-frequenti/documento-informatico. Consultato il 27/02/2023

conservazione dei documenti informatici, alfine di fruire delle informazioni di interesse;

- Conservatore è il soggetto pubblico o privato che svolge attività di conservazione dei documenti informatici;
- Responsabile del servizio di conservazione soggetto che coordina il processo di conservazione all'interno del conservatore, in possesso dei requisiti professionali individuati da AGID nel *Regolamento sulla fornitura dei servizi di conservazione dei documenti informatici.*

Se tale elenco chiarisce il ruolo di ciascun soggetto nel processo di conservazione e valorizza la figura del responsabile della conservazione, che deve essere in possesso di specifiche competenze archivistiche, giuridiche e informatiche, comporta anche alcune perplessità in merito alla loro individuazione. Infatti, i servizi di conservazione erogati dai Conservatori sono spesso incentrati su soluzioni tecnologiche automatizzate per il versamento dei documenti (per esempio SFTP o Web Services). Ciò da un lato ottimizza tale operazione di versamento nel sistema di conservazione dall'altro implica la sovrapposizione dei ruoli complicando l'individuazione della figura del Produttore dei PdV.

Tra le disposizioni dettate dalle Linee guida figura anche l'obbligo tanto per le amministrazioni pubbliche quanto per i privati di formare i documenti secondo le modalità previste dal par. 2.1.1. Le modalità di formazione non differiscono da quelle previste dalle precedenti regole tecniche. Tuttavia, l'aver ribadito da parte del legislatore la possibilità di formare documenti giuridicamente rilevanti anche mediante la c) *memorizzazione su supporto informatico in formato digitale delle informazioni risultanti da transazioni o processi informatici o dalla presentazione telematica di dati attraverso moduli o formulari resi disponibili all'utente* o la *d) generazione o raggruppamento anche in via automatica di un insieme di dati o registrazioni, provenienti da una o più banche dati, anche appartenenti a più soggetti interoperanti, secondo una struttura logica predeterminata e memorizzata in forma statica*[60] ha richiesto alla comunità degli addetti ai lavori un'attenta riflessione sulle metodologie e tecnologie necessarie per garantire la conservazione delle basi di dati attraverso le quali sono formate i documenti al fine di mantenere nel tempo i requisiti di autenticità, affidabilità e leggibilità. Infatti, nello stesso paragrafo il legislatore indica solo le operazioni necessarie per assicurare la immodificabilità di un documento informatico e nel caso specifico della lettera c) e d) questa è garantita o dalla:

- apposizione di una firma elettronica qualificata, di una firma digitale o di un sigillo elettronico qualificato o firma elettronica avanzata;

[60] Si veda il paragrafo 2.1.1. "*Formazione del documento informatico*" in "*Linee Guida sulla formazione, gestione e conservazione dei documenti informatici*"

- registrazione nei log di sistema dell'esito dell'operazione di formazione del documento informatico, compresa l'applicazione di misure per la protezione dell'integrità delle basi di dati e per la produzione e conservazione dei log di sistema;
- produzione di una estrazione statica dei dati e il trasferimento della stessa nel sistema di conservazione.

Tale interesse trova un sostanziale riscontro nella istituzione del sottogruppo di lavoro 2 nell'ambito del Gruppo di lavoro sui Poli di Conservazione[61]. La tematica della conservazione delle basi di dati costituisce una sfida ancora aperta anche nel contesto internazionale che non può essere ridotta alla realizzazione di una semplice procedura informatica di *back-up* né tantomeno all'estrazione e il riversamento della struttura e dei valori della base di dati, ma deve tener poter preservare il complesso informativo necessario per garantire il rispetto degli elementi archivistici essenziali e il valore probatorio dei documenti. Tra le esperienze internazionali riservate alla conservazione delle basi di dati figura il *First International Workshop on Database Preservation (PresDB'07)* organizzato dall'UK *Digital Curation Centre* (DCC) presso il *National e-Science Centre* di Edimburgo il 23 marzo 2007. Obiettivo dell'evento è identificare le possibili criticità tecnologiche, economiche e legali legate alla conservazione delle basi di dati. Le ragioni dell'obiettivo sono chiarite da Cristophides et al. che evidenzia come «*while substantial investment has been made in archiving and preserving conventional forms of these objects, such as documents, images and numerical data in some file format, the need to preserve entire database has only recently emerged*»[62]. Per offrire una panoramica complessiva del tema, l'evento è organizzato in una duplice prospettiva informatica e archivistica. Una seconda esperienza è il progetto RODA (Repository of Authentic Digital Objects)[63] ideato e avviato nel 2006 dal *National Archives of Portugal* (Directorate-General of the Portoguese Archives). Scopo del progetto «*is to provide technical solution to digital preservation*»[64] alle pubbliche amministrazioni portoghesi. Ciò che ne deriva è la modellazione di un flusso di processi per la conservazione degli oggetti digitali incentrato sull'uso di standard internazionali quali l'ISO 14721:2012 Open Archival Information System (OAIS)[65], il Metadata Encoding and Transmission Standard (METS)[66], l'Encoded Archival Description (EAD)[67],

[61] Agenzia per l'Italia Digitale, "Vademecum per l'implementazione delle linee guida sulla formazione, gestione e conservazione dei documenti informatici, 2022, link: https://www.agid.gov.it/it/linee-guida. Consultato il 27/02/2023

[62] CHRISTOPHIDES, Vassilis, BUNEMAN, Peter, "Report on the First International Workshop on Database Preservation (PresDB'07)", in SIGMOD Record, Vol. 36, No. 3, 2007

[63] Università di Porto, "RODA project", link: http://www.roda-community.org. Consultato il 27/02/2023;

[64] RAMALHO, Jose Carlos, FERREIRA, Miguel, FARIA, Luis, CASTRO, Rui, "Relational Database Preservation Through XML Modelling", in Extreme Markup Languages 2007, Montréal, 2007

[65] ISO 14721:2012 Space data and information transfer systems — Open archival information system (OAIS) — Reference model

[66] The Library of Congress, "METS. Metadata Encoding & Trasmission Standard Official Web site", Ivi

[67] The Library of Congress, "<ead>. Encoded Archival Description", Ivi

il Dublin Core (DC)[68], il PREMIS (Preservation Metadata)[69] e di tecnologie open-source. In particolare, tale progetto ha comportato lo sviluppo del Database Preservation Toolkit[70] (DBPTK) per supportare la migrazione tra i formati delle basi di dati relazionali ai fini della conservazione all'interno del repository RODA.

Infine, l'esperienza più rilevante è il rilascio del formato SIARD nel 2004 (Software Indipendent Archiving of Relational Database) e del relativo software per estrarre dati, metadati e relazioni da una base di dati relazionale per poterli migrare all'interno di una nuova base di dati strutturati secondo la sintassi del formato SIARD, basata sulle specifiche XML, SQL:2008, UNICODE e ZIP64[71]. Tuttavia, tali esperienze non hanno ancora raggiunto una maturità definitiva poiché tendono a valorizzare gli strumenti tecnologici e a ignorare gli elementi archivistici essenziali con ricadute negative in merito all'autenticità e all'affidabilità dei documenti. Per esempio, da una gap analysis condotta in merito all'uso del formato SIARD, ampiamente adoperato nella maggior parte delle esperienze di conservazione delle basi di dati relazionali (tra cui l'archivio RODA), è stato osservato come tale formato non garantisce l'estrazione delle *primary keys* necessarie a definire le relazioni tra le tabelle. Ciò determina l'impossibilità di aggregare i dati, generare i documenti, ricostruire il vincolo archivistico e il contesto di produzione. La comunità scientifica ha avanzato diverse proposte per mitigare tali gap, tra cui anche modificare direttamente il codice sorgente per integrare i moduli dedicati all'estrazione delle informazioni indicate. Tuttavia, attualmente il software non permette direttamente l'aggiornamento del codice.

Un ruolo centrale ai fini della conservazione dei documenti informatici è ricoperto dall'allegato 2 delle Linee guida inerente i formati di file e riversamento[72]. Tale allegato fornisce i formati di file consigliati e non consigliati ai fini della conservazione. La scelta del formato è un aspetto cruciale ai fini della conservazione fin dal momento della formazione del documento che deve essere incentrata sulla valutazione dei criteri indicati dall'indice della interoperabilità del medesimo allegato. Tale indice include la valutazione complessiva del formato il cui punteggio costituisce la somma delle caratteristiche dello stesso. Da tale analisi si desume che un formato dovrebbe essere *de iure, aperto, non proprietario, estendibile, completamente robusto, indipendente dal dispositivo, retrocompatibile e possieda un alto livello di metadati.* Tale formato deve essere adottato in rapporto all'oggetto digitale e per, supportare i

[68] OCLC, "Dublin Core (Dublin Core)", link: https://www.dublincore.org/ . Consultato il 27/02/2023

[69] The Library of Congress, "PREMIS. Preservation Metadata Maintenance Activity, Ivi

[70] RAMALHO, Jose Carlos, FARIA, Luis, SILVA, Hélder, Coutada, Miguel, "DB-Preservation Toolkit in GitHub), http://keeps.github.io/db-preservation-toolkit/ attualmente migrato su: https://github.com/keeps/dbptk-developer. Consultato il 27/02/2023

[71] Archivio Federale Svizzero, "SIARD Suite", link: https://www.bar.admin.ch/bar/it/home/archiviazione/strumenti/siard-suite.html. Consultato il 27/02/2023

[72] Agenzia per l'Italia Digitale, "Allegato 2. Formati di file e riversamento" in Linee Guida sulla formazione, gestione e conservazione dei documenti informatici, Roma, 2021;

soggetti nell'adempimento della scelta, il legislatore ha suddiviso tali oggetti in diverse tipologie.

L'estensione anche ai formati non consigliati per la conservazione e l'interoperabilità è motivata dal dall'ampia diffusione nella pubblica amministrazione. Per mitigare il rischio dell'obsolescenza tecnologica, l'allegato include le specifiche relative al riversamento di un documento informatico in un formato di file diverso. Tuttavia, tale trasferimento determina la generazione di una sequenza binaria differente rispetto al documento originale e quindi la formazione di una copia informatica di un documento informatico. Tale asserzione è ripresa anche dal legislatore nel par. 2.3. delle linee guida in cui il legislatore stabilisce che "la copia di un documento informatico è un documento il cui contenuto è il medesimo dell'originale ma con una diversa evidenza informatica rispetto al documento da cui è tratto, come quando si trasforma un documento con estensione ".doc" in un documento ".pdf"[73], il cui valore probatorio è il medesimo all'originale se la conformità non è espressamente disconosciuta o se è attestata da un pubblico ufficiale autorizzato in tutte le sue componenti.

Un ulteriore adempimento ai fini della conservazione concerne la generazione e l'associazione dei metadati. In particolare, l'allegato 5 indica il complesso di metadati suddividendoli in metadati del documento informatico, metadati del documento amministrativo informatico e metadati delle aggregazioni documentali informatiche. Tale allegato introduce importati novità rispetto alle precedenti regole tecniche. Dal punto di vista formale, non è previsto il concetto di set minimo di metadati poiché sono suddivisi in obbligatori e non obbligatori. Inoltre, il numero è incrementato a 18 metadati per il documento informatico e il documento amministrativo informatico (con alcune differenze, tra cui la segnatura e l'obbligatorietà di alcuni metadati da associare al documento amministrativo informatico rispetto al documento informatico) e 15 per le aggregazioni documentali informatiche. Dal punto di vista strutturale, gli schemi di metadati indicano per ciascun metadato l'*element :xsd*, il *Campo*, i relativi *Sottocampi,* i *Valori ammessi,* il *Tipo dato,* l'*Obbligatorietà e* l'eventuale *Nuova definizione.* L'adozione del nuovo schema di metadati da un lato ha arricchito le informazioni a corredo dei documenti per garantire tra gli altri il recupero e la leggibilità nel tempo, dall'altro ha sollevato diverse criticità sia da parte dei Titolari dell'oggetto di conservazione sia da parte dei Conservatori. Tali criticità includono tanto aspetti prettamente archivistici che tecnologici suddivisibili concettualmente in generali e particolari. Nel dettaglio, tra le criticità a livello generale figura l'utilizzo di tecnologie non adeguate a supportare schemi di metadati annidati la cui modifica comporta l'investimento anche di risorse economiche di cui non dispongono i soggetti produttori. La mancata adozione di uno schema annidato determina l'adozione di schemi piatti e,

[73] Par. 2.3 "*Duplicati, copie ed estratti informatici di documenti informatici*" in *Linee Guida sulla formazione, gestione e conservazione dei documenti informatici* di AgID

sebbene AgID abbia chiarito che la struttura e il formato dei metadati non sono obbligatori, ciò comporta l'adozione di soluzioni che determinano ricadute negative in termini di elaborazione delle informazioni e del relativo recupero. Inoltre, i Conservatori hanno previsto soluzioni utili a garantire l'immodificabilità dei valori dei metadati al momento del versamento ma ciò determina la difficoltà di compilazione dei metadati di versionamento (*VersioneDelDocumento*) nel sistema di conservazione. Tra le criticità a livello particolare figurano per esempio la genericità della descrizione fornita per ciascun metadato. Per fornire un supporto, AgID ha emanato il documento "*I metadati del documento informatico di natura fiscale e contabile*" volto a chiarire la corretta compilazione dei metadati delle suddette tipologie documentali.

Tuttavia, la mancata esaustività delle tipologie documentali lascia ancora spazio a perplessità nel caso di associazione dei metadati a documenti non di natura fiscale e contabile (per esempio amministrativa, tra cui la corrispondenza).

Inoltre, l'obbligatoria indicazione per le amministrazioni pubbliche dei metadati relativi alla classificazione (*IndiceDiConservazione* e *Descrizione*) ha sollevato la mancata stesura del Titolario di classificazione la cui redazione era obbligatoria sin dalla emanazione del testo unico DPR 445/2000[74], mentre la non obbligatorietà dei metadati utili a rappresentare i tempi di conservazione (*TempoDiConservazione*) potrebbe comportare ricadute negative in termini di selezione e scarto della documentazione.

Infine, l'ultimo, in ordine numerico, degli allegati che abroga e sostituisce la circolare n. 60 del 23 gennaio 2013 dell'AgID in materia di "Formato e definizione dei tipi di informazioni minime ed accessorie associate ai messaggi scambiati tra le Pubbliche Amministrazioni" è l'allegato 6 "Comunicazione tra AOO di documenti amministrativi protocollati". È evidente che quest ultimo allegato non riveste particolare interesse per il tema di ricerca trattato nel presente lavoro di tesi dal momento che pochi tra i documenti della ricerca riceveranno una protocollazione e/o potranno essere scambiati all'interno di AOO.

L'allegato 6 descrive le specifiche tecniche obbligatorie relative alle informazioni che costituiscono la segnatura di protocollo associata ai documenti amministrativi informatici protocollati al fine di favorire le interoperabilità tra Aree Organizzative Omogenee (AOO). Tali informazioni devono essere strutturate secondo le specifiche del formato XML affinché prevedano obbligatoriamente la sezione relativa all'intestazione, che include i dati identificativi e le informazioni fondamentali del messaggio, quella relative alla *signature*, che permette la sottoscrizione della segnatura e opzionalmente le informazioni relative al messaggio di protocollo ricevuto e quelle riguardanti il contenuto del messaggio. Una delle modalità di inoltro del messaggio di protocollo adoperata dalle pubbliche amministrazioni è incentrato sull'uso della posta elettronica

[74] In particolare, si veda l'art. 52 "*Il sistema di gestione informatica dei documenti*" del *Decreto del Presidente della Repubblica 28 dicembre 2000, n. 445 Testo unico delle disposizioni legislative e regolamentari in materia di documentazione amministrativa* (G.U. n. 42 del 20 febbraio 2001, s.o. 30/L)

certificata (PEC) normata dal DPR n. 68/2005[75] e dall'art. 80 del CAD che ne riconosce il medesimo valore legale di una raccomandata con ricevuta di ritorno.

La complessità delle disposizioni e le indicazioni derivanti dalle specifiche tecniche inducono a pensare che gli strumenti normativi in uso siano sufficienti per garantire la conservazione dei documenti digitali e mantenere i requisiti di autenticità, affidabilità, integrità e accessibilità nel tempo. Tuttavia, l'adozione delle norme in domini applicativi caratterizzati da un complesso documentale eterogeneo solleva alcune perplessità. In particolare, le Linee Guida disciplinano dettagliatamente la gestione documentale con particolare attenzione ai documenti amministrativi protocollati ma ignorano il complesso di tipologie documentali che potrebbero essere prodotte da una pubblica amministrazione. Per esempio, i prodotti della ricerca includono un complesso di informazioni prodotte durante l'intero ciclo di vita della ricerca che attestano le risorse investite, le tematiche trattate, i tempi di realizzazione, i soggetti coinvolti, le metodologie utilizzate e i risultati ottenuti. Una prima reazione internazionale alla complessa questione della gestione dei prodotti della ricerca[76] è stata la nascita del movimento *Open Access*[77] durante la *Budapest Open Access Initiative* (BOAI) nel 2002 e la definizione di strategie alternative per la comunicazione scientifica note come green-road e gold-road[78]. Alla BOAI si sono succeduti altri eventi a sostegno dell'Open Access, tra cui *Bethesda Statement* e *Berlin Declaration* e la *Dichiarazione di Messina*. Anche l'Unione Europea ha sollecitato gli stati membri nella diffusione dei risultati della ricerca scientifica. A titolo esemplificativo basti pensare ai recenti cospicui finanziamenti investiti per assicurare l'interoperabilità tra centri di ricerche in merito ai dati ottenuti dalle ricerche sul COVID-19 o ai noti programmi *Horizon 2020* e successivamente il quadro *Horizon Europe*[79] che promuovono entrambi le politiche di Open Science anche tramite il progetto *OpenAire*[80].

In entrambi gli scenari, i repository devono essere implementati su policy adeguate che assicurino la gestione delle informazioni durante l'intero ciclo di vita della ricerca fino alla conservazione e che tutelino l'accesso alle informazioni secondo i diritti previsti. In particolare, l'adozione formati aperti e interoperabili e di schemi di metadati incentrati su standard internazionali determina la possibilità, per i ricercatori di diverse discipline, di accedere e riutilizzare tale patrimonio informativo per aumentare la

[75] Presidente della Repubblica, "Decreto del Presidente della Repubblica recante il Regolamento recante disposizioni per l'utilizzo della posta elettronica certificata a norma dell'articolo 27 della legge 16 gennaio 2003, n. 3", 11 febbraio 2005, n. 68 (G.U. n.97 del 28-04-2005)

[76] Il movimento *OpenAccess* nasce per fronteggiare l'aumento dei prezzi imposti dalle riviste scientifiche (anche noto come "*Crisi dei prezzi*") e per fornire soluzioni tecnologiche al fine di migliorare la conservazione e aumentare l'interoperabilità tra repository

[77] CASSELLA, Maria, *Open Access e comunicazione scientifica* … cit.

[78] GUERRINI, Mauro, *Gli Archivi Istituzionali* … cit.

[79] Consiglio dell'Unione Europea, "Proposta di decisione del consiglio relativa all'istituzione del programma specifico di attuazione di Orizzonte Europa - il programma quadro di ricerca e innovazione", link: https://data.consilium.europa.eu/doc/document/ST-8550-2019-INIT/it/pdf. Consultato il 27/02/2023

[80] Unione Europea, "OpenAIRE", link: https://www.openaire.eu/. Consultato il 27/02/2023;

disseminazione della conoscenza, migliorare il processo di knowledge discovery, favorire l'analisi interdisciplinare dei fenomeni oggetto d'indagine, ottimizzare gli investimenti pubblici e aprire al territorio le scoperte scientifiche anche in ottica collaborativa tra le realtà imprenditoriali e i centri di ricerca.

Tuttavia, le università e i centri di ricerca propongono di fatto set minimi di metadati che non risultano utili a garantire il controllo del valore informativo e la possibilità di accesso alle risorse nel tempo. Per esempio, l'adozione di schemi con un alto livello di astrazione causa una descrizione poco significativa delle risorse mentre la definizione arbitraria dei metadati limita l'interoperabilità tra repository. Pertanto, diviene inderogabile ripensare le attuali metodologie di *digital-transformation* in direzione di una riconsiderazione delle procedure relative alla conservazione dei documenti digitali e in particolare dei prodotti della ricerca. In particolare, tale ripensamento deve essere incentrato sulla presa di consapevolezza che il digitale non rappresenta solo uno strumento estemporaneo da adottare per adempiere ad obblighi normativi, ma costituisce un elemento di semplificazione cruciale se valorizzato adeguatamente attraverso l'adozione di procedure incentrate elevati requisiti di qualità e sicurezza, formati aperti e interoperabili (come XML o RDF utilizzato per i Knowledge graph) e standard di metadati internazionali per supportare sia il recupero delle risorse sia la più rapida diffusione dei linked open data nel semantic web incremento la disseminazione della conoscenza scientifica.

La conservazione digitale è un requisito essenziale per qualunque organizzazione che intende intraprendere la transizione di documenti giuridicamente rilevanti dal tradizionale supporto cartaceo a quello informatico. Tale requisito è finalizzato a garantire l'autenticità, l'affidabilità, il recupero e la leggibilità dei documenti digitali. Per supportare i Paesi nel processo transizionale, l'Unione Europea ha varato una serie di misure, tra cui il progetto "Europa 2020" di cui l'Agenda Digitale Europea (ADE)[81] costituisce una delle sette iniziative faro emanata dalla Commissione Europea. In particolare, l'ADE individua sette criticità principali da superare attraverso la valorizzazione delle tecnologie dell'informazione e della comunicazione (*Information and Communication Technologies, ICT*):

- *Frammentazione dei mercati digitali*, eliminare le barriere normative e agevolare le fatturazioni e i pagamenti elettronici, la risoluzione delle controversie e rafforzando la fiducia dei consumatori attraverso un quadro normativo che crei un mercato unico nel settore delle telecomunicazioni;
- *Mancanza di interoperabilità,* abbattere le carenze in materia di definizione degli standard, appalti pubblici e coordinamento tra amministrazioni pubbliche che impediscono ai servizi e ai dispositivi digitali utilizzati dai cittadini europei di

[81] Commissione Europea, "Comunicazione della commissione al parlamento europeo, al consiglio, al comitato economico e sociale europeo e al comitato delle regioni. Un'agenda digitale europea", Bruxelles, 19 maggio 2010

funzionare insieme prevedendo l'uso di componenti e applicazioni che siano interoperabili e si basino su standard comuni e piattaforme aperte

- *Aumento della criminalità informatica e rischio di un calo della fiducia nelle reti,* combattere le nuove forme di criminalità informatica tra cui il furto di identità e a i dati personali che minano la tutela dei diritti fondamentali degli europei, tra cui la riservatezza;

- *Mancanza di investimenti nelle reti,* migliorare la velocità della rete internet attraverso investimenti nelle tecnologie sia fisse che senza fili che migliori lo scambio dei dati, tra cui la banda larga;

- *Impegno insufficiente nella ricerca e nell'innovazione,* fare leva sul talento dei ricercatori per creare un clima di innovazione nel quale le aziende europee di qualunque dimensione che operano nel settore delle TIC possano mettere a punto prodotti eccellenti in grado di generare una domanda;

- *Mancanza di alfabetizzazione digitale e competenze informatiche,* combattere l'analfabetismo digitale per evitare che i cittadini europei siano esclusi dalla società e dall'economia digitale con ricadute negative anche in termini economici attraverso iniziative dirette da parte degli Stati membri e delle altre parti interessate;

- *Opportunità mancate nella risposta ai problemi della società,* superamento di problemi pressanti per la comunità, come ad esempio i cambiamenti climatici e le altre pressioni sull'ambiente, l'invecchiamento demografico e i costi sanitari crescenti, lo sviluppo di servizi pubblici più efficienti e l'integrazione delle persone con disabilità e la digitalizzazione del patrimonio culturale europeo per metterlo a disposizione della generazione attuale e di quelle future.

Le soluzioni adottate dall'Unione Europea per superare gli ostacoli individuati dall'ADE si sono tradotte nel tempo in misure normative, tra cui, il regolamento n. 910/2014 in materia di identificazione elettronica e servizi fiduciari per le transazioni elettroniche nel mercato interno (electronic IDentification Authentication and Signature - eIDAS) che include le diverse tipologie di firme elettroniche[82] e il regolamento n. 679/2016 relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati (General Data Protection Regulation – GDPR)[83]. A seguito della prima ADE, l'Unione Europea ha

[82] Parlamento Europeo, "Regolamento (UE) n. 910/2014 del Parlamento europeo e del consiglio del 23 luglio 2014 in materia di identificazione elettronica e servizi fiduciari per le transazioni elettroniche nel mercato interno e che abroga la direttiva 1999/93/ce", Bruxelles, 28 agosto 2014

[83] Parlamento Europeo, "Regolamento (UE) n. 2016/679 del Parlamento europeo e del consiglio del 27 aprile 2016 relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati e che abroga la direttiva 95/46/CE (regolamento generale sulla protezione dei dati)", Bruxelles, 27 aprile 2016

adottato la seconda ADE che analizza i cambiamenti introdotti dalle tecnologie digitali, il ruolo essenziale svolto dai servizi, dai mercati digitali e le nuove ambizioni dell'UE in campo tecnologico e geopolitico. In particolare, la Commissione Europea intende intraprendere entro il 2030 le misure previste in due documenti strategici: "*Plasmare il futuro digitale dell'Europa*"[84] e «*Il decennio digitale europeo*»[85]. Tra le priorità figurano lo sviluppo della computazione quantistica, una strategia in materia di blockchain e una politica commerciale basata sulla blockchain, l'intelligenza artificiale antropocentrica e affidabile, i semiconduttori (normativa europea sui semiconduttori), la sovranità digitale, la cybersicurezza, la connettività Gigabit, il 5G e il 6G, gli spazi e le infrastrutture europee dei dati, nonché la definizione di norme tecnologiche globali. Per attuare tali obiettivi l'Unione Europea ha stanziato un budget di circa 7.5 miliardi di euro attraverso il *Programma Digitale Europeo* (DIGITAL)[86] grazie al quadro finanziario pluriennale 2021-2027. In particolare, il DIGITAL intende sostenere progetti in cinque aree chiave di capacità: supercalcolo, intelligenza artificiale, cybersicurezza, competenze digitali avanzate e garantendo un ampio uso delle tecnologie digitali in tutta l'economia e la società, anche attraverso il digitale Hub dell'Innovazione. Uno dei primi risultati prodotti dall'investimento delle risorse per la valorizzazione dell'Intelligenza Artificiale è il "*Libro Bianco sull'intelligenza artificiale - Un approccio europeo all'eccellenza e alla fiducia*" pubblicato nel 2020 dalla Commissione Europea. Obiettivo del libro bianco è presentare le opzioni strategiche per consentire uno sviluppo sicuro e affidabile di un'Intelligenza Artificiale antropocentrica all'interno del territorio giuridico europeo per conseguire un "ecosistema di eccellenza" e "un ecosistema di fiducia".

La politica europea in materia di sviluppo del mercato unico punta molto sulla realizzazione di un ecosistema digitale in cui vi sia piena realizzazione dei principi di interoperabilità, di affidabilità e di accesso ai documenti. Viene così definito uno scenario nel quale la conservazione trova un suo ruolo strategico fondamentale anche in relazione allo sviluppo dei territori e del tessuto produttivo. Sulla base di tali

---

Altre misure riguardano:

- Parlamento Europeo, "*Regolamento (UE) n. 2018/1807 sulla libera circolazione dei dati non personali che consente alle imprese e alle amministrazioni pubbliche di archiviare e trattare i dati non personali ovunque scelgano di farlo nell'UE"*, Bruxelles, 14 novembre 2018;
- Parlamento Europeo, "*Regolamento sulla cybersicurezza (UE) n. 881/2019, che rafforza l'Agenzia dell'Unione europea per la cybersicurezza (ENISA) e istituisce un quadro per la certificazione della cybersicurezza di prodotti e servizi"*, Bruxelles, 17 aprile 2019;
- Parlamento Europeo, "*Direttiva sull'apertura dei dati (UE) n. 2019/2014 che stabilisce norme comuni per un mercato europeo per i dati in possesso del governo*", Bruxelles, 20 giugno 2019.

[84] Parlamento Europeo, "L'Agenda Digitale Europea", link: https://www.europarl.europa.eu/factsheets/it/sheet/64/un-agenda-digitale-europea . Consultato il 27/02/2023

[85] Commissione Europea, "Decennio Digitale Europeo: Obiettivi digitali per il 2030", link: https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/europes-digital-decade-digital-targets-2030_it. Consultato il 27/02/2023;

[86] Commissione Europea, "Digital Europe Programme", link: https://ec.europa.eu/info/funding-tenders/find-funding/eu-funding-programmes/digital-europe-programme_it. Consultato il 27/02/2023;

considerazioni l'Italia ha inteso, per prima tra i paesi dell'Unione, investire nella costruzione di un modello di conservazione che ricorre anche a forme di standardizzazione che pur essendo nate nel contesto nazionale aspirano ad una possibile applicazione estesa a tutti i paesi dell'Unione. Tra i modelli implementati in tema di conservazione digitale vale la pena di ricordare lo standard nazionale UNI 11386 "Supporto all'Interoperabilità nella Conservazione e nel Recupero degli Oggetti digitali (SInCRO)"[87]. La norma tecnica è stata pubblicata nel maggio 2020 e successivamente richiamata dal legislatore nelle *Linee guida sulla formazione, gestione e conservazione dei documenti informatici*. La versione attualmente in vigore ha apportato numerose modifiche alla precedente versione pubblicata nel 2010, tra cui la normalizzazione linguistica degli elementi esclusivamente in inglese, l'ampliamento degli elementi e la contestualizza del loro utilizzo nel processo di conservazione digitale dei documenti informatici.

Tale, norma definisce gli elementi informativi per la creazione dell'Indice di Conservazione (Preservation Index) da parte del responsabile della conservazione interno al titolare dell'oggetto di conservazione (soggetto produttore) o dal responsabile del servizio di conservazione interno al conservatore in caso di affidamento all'esterno del servizio di conservazione. In particolare, lo schema è elaborato secondo il linguaggio XML ed è basato su una struttura gerarchica al cui apice vede l'elemento radice *PIndex* e i relativi quattro elementi: ***SelfDescription, PVolume, FileGroup*** e ***Process*** e l'elemento ***Agent*** indipendente dalla radice PIndex.

L'indice di conservazione definito dallo standard UNI SiNCRO 11386 permette di raggiungere un primo grado di interoperabilità sintattica e semantica tra i sistemi di conservazione, non senza alcune ombre che probabilmente troveranno soluzione nelle future versioni dello standard stesso. Infatti, l'applicazione nei contesti di interoperabilità ha già evidenziato alcuni elementi di criticità quali, ad esempio, la non ripetibilità della classe di informazioni MoreInfo all'interno di ciascuna sezione.

Riepilogando, i metadati rivestono un ruolo cruciale nella conservazione digitale di qualunque risorsa informativa. In particolare, la definizione di uno schema adeguatamente strutturato sulla base di standard internazionali di riferimento documenta ogni fase del ciclo di vita della risorsa assicurandone il recupero, la riproducibilità e la leggibilità nel tempo, mentre l'uso di formati aperti (per esempio XML) permette l'interoperabilità tra i sistemi e l'indipendenza tecnologica da risorse hardware e software, con ricadute positive sia in termini di efficacia, di efficienza e di economicità. Nel dettaglio, i metadati di un archivio digitale devono descrivere tanto le informazioni del singolo documento informatico quanto quelle sull'intero contesto di produzione per assicurare valore probatorio alle rappresentazioni informatiche. Per esempio, è necessario che la conservazione di archivi digitali contenenti le testimonianze amministrative e scientifiche sia incentrata sull'uso di metadati che

[87] UNI 11386:2020 "Supporto all'Interoperabilità nella Conservazione e nel Recupero degli Oggetti digitali. SInCRO"

permettano di relazionare le informazioni per ricostruire il vincolo archivistico tra i documenti amministrativi e i prodotti della ricerca afferenti al medesimo procedimento amministrativo e su metadati che identifichino chiaramente gli agenti e gli eventi in rapporto ai diritti posseduti e che siano incentrati sull'uso di formati come XML al fine di garantire l'interoperabilità tra i sistemi e la disseminazione delle scoperte scientifiche in rapporto alle licenze d'uso utilizzate.

In tal senso, relazionare le informazioni costituisce l'assetto centrale per la generazione di conoscenza. In particolare, come teorizzato da Russell Lincoln Ackoff[88] esiste una relazione intrinseca tra dato, informazione, conoscenza, comprensione e saggezza le cui relazioni determinano la nascita della cultura così come espressamente formulata nella "*Piramide di Ackoff*".

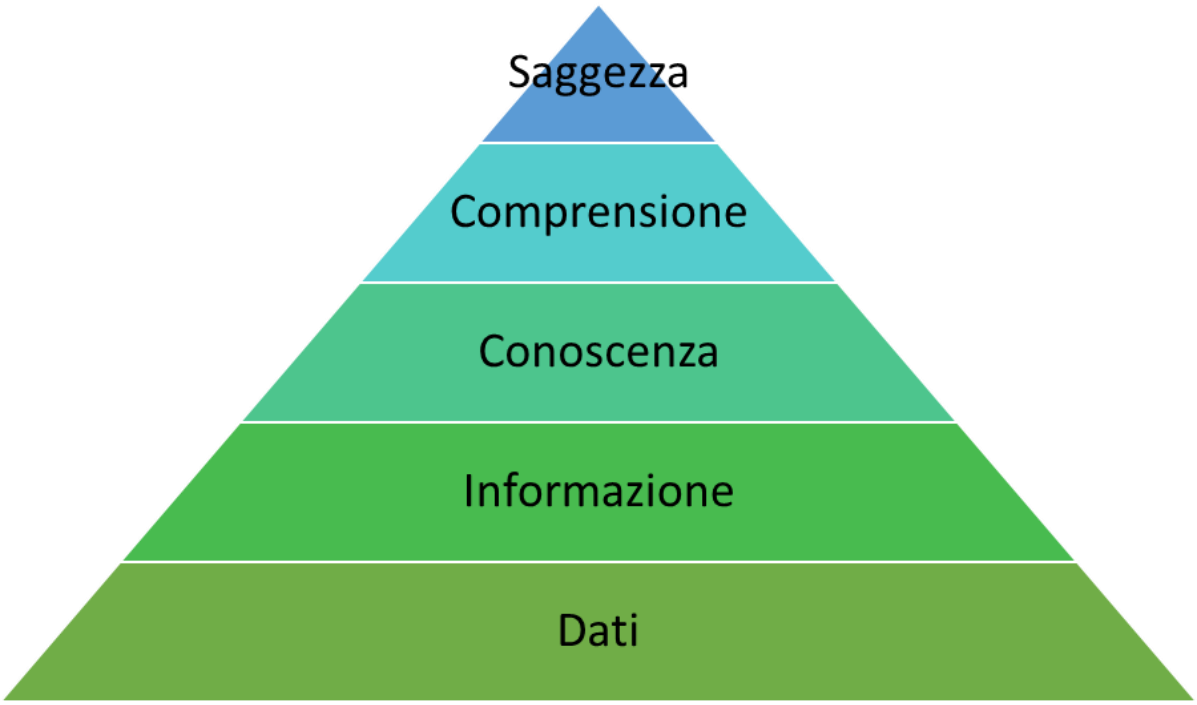

Figura 6. La piramide di Ackoff

Nel contesto della trasformazione digitale tali relazioni assumono un ruolo centrale per valorizzare le risorse investite da qualunque organizzazione. In tale contesto, il dato può essere considerato il valore contenuto all'interno di una tabella prevista in una base di dati rappresentato da una stringa alfanumerica. Per esempio, "Mario" potrebbe essere un dato valoriale contenuto all'interno di una cella. La relazione tra l'attributo che indica la tipologia di valore e il dato inserito all'interno di una cella costituisce una informazione. Nel caso preso in esame, la relazione tra l'attributo "Nome" e il valore contenuto nella "Mario" genera una informazione relativa a un "Nome" di tipo "Mario". La relazione significativa tra le informazioni genera conoscenza. Il termine significativo rimanda inevitabilmente a modelli interpretativi

[88] ACKOFF, Russell Lincoln, "Ackoff From Data to Wisdom", in *Journal of Applied System Analysis*, 1989;

utili a supportare l'analisi delle relazioni tra le informazioni per la scoperta della conoscenza. Per esempio, la relazione tra gli attributi "Nome" e "Cognome" genera una conoscenza in merito all'esistenza di un "Nome" di tipo "Mario" e di un "Cognome" di tipo "Rossi". In tal senso, maggiori saranno le relazioni tra le informazioni maggiormente complessa sarà la conoscenza di quel dominio. Per affinare la conoscenza di un determinato dominio sono stati sviluppati algoritmi di intelligenza artificiale incentrati su tecniche machine e deep learning in grado di analizzare anche grandi moli di dati per estrarne modelli (patterns) che definiscono la conoscenza di uno specifico dominio e supportano i decisori nelle strategie future (*decision making*). In tal senso, l'interpretazione dei modelli genera conoscenza che a sua volta determina la comprensione (*understanding)* di un dominio che richiede inevitabilmente la scelta da adoperare in rapporto all'obiettivo prefissato (per esempio la definizione di un piano strategico economico che preveda l'investimento o meno di risorse per l'assunzione di figure competenti o l'implementazione di specifiche tecnologie).

1.4. Estrazione automatica di metadati

Uno dei possibili approcci alle criticità di rappresentazione dei prodotti della ricerca è quello di estrazione automatica di metadati. L'obiettivo finale di tale procedura consiste nell'ottimizzare la descrizione delle risorse, supportare l'estrazione di conoscenza e assicurare la conservazione digitale. In particolare, il processo di estrazione ricorre a framework implementati su architetture multilivello, in cui ogni livello è costituito da uno o più tool logicamente connessi, per cui l'output prodotto dall'analisi di un tool costituisce l'input del tool successivo, fino a ottenere il risultato finale contenente i metadati estratti dalla risorsa. La scelta dei tool dipende da diversi fattori in rapporto all'obiettivo prefissato come il dominio applicativo, il contesto tecnologico, i formati utilizzati e i metadati che si desidera estrarre. È possibile, ad esempio, distinguere tra sistemi di estrazione automatica distribuita di metadati e sistemi di estrazione automatica centralizzata. L'implementazione di sistemi di estrazione distribuita di metadati è riconducibile all'esigenza di gestire in modo efficace ed efficiente l'esponenziale incremento di dati (da cui il termine *Big Data*) provenienti da diverse fonti (*Internet of Things)* interconnesse attraverso tecnologie wired e wireless[89] con differenti formati. Tale esigenza trova nelle basi di dati non-relazionali una prima risposta immediata determinata dalla impossibilità dei tradizionali DBMS[90] relazionali di manipolare (ricevere, organizzare, preparare, trasformare, determinare le dipendenze K→V (Key→Value), acquisire e reperire) i dati non conoscibili a priori. Nel contesto scientifico, per esempio, «*data lakes follow the same idea: data is extracted from the sources and is stored in its original structure in a repository which is often based on Hadoop or NoSQL database systems*»[91]. Estrarre automaticamente metadati in un contesto distribuito è essenziale per supportare i provider nelle procedure di *data analysis* per la customizzazione di policy di cloud computing, la gestione delle risorse e la valorizzazione del patrimonio conoscitivo. In tal senso, il valore dei dati «*is obtained by analysing Big Data and extracting from them hidden patterns, trends and knowledge models by using smart data analysis algorithms and techniques*»[92]. Sebbene l'estrazione automatica di metadati da ambiente distribuito non costituisca l'oggetto del presente lavoro vale comunque la pena citare alcune soluzioni. Un esempio di framework di estrazione automatica di metadati implementabile sia localmente sia on-demand è *Skluma*. In particolare, *Skluma* è in grado di estrarre automaticamente i metadati sui topic da dati testuali non strutturati e sulla conoscenza tacita delle immagini trattati all'interno di file system condiviso in repository di dati scientifici. Dal

[89] Per esempio: Ethernet, WI-FI, Bluetooth, ZigBee, GSM, GPRS, GPS

[90] Per maggiori dettagli sui *Datalakes* si suggerisce: MATHIS, Christian, "Data Lakes", in *Springer*, 2017;

[91] QUIX, Christoph, HAI, Rihan, VATOV, Ivan, "Metadata Extraction and Management in Data Lakes With GEMMS", in Complex Systems Informatics and Modeling Quarterly, 2017. Pages 67–83;

[92] ELSHAVWI, Radwa, SAKR, Sherif, TALIA, Domenico, TRUNFIO, Paolo, "Big Data Systems Meet Machine Learning Challenges: Towards Big Data Science as a Service", in Journal of Big Data Research, Elsevier, 2018;

punto di vista strutturale, l'architettura (figura 1) è basata sull'iniziale elaborazione e indicizzazione dei file system da parte di un *Crawler*. Successivamente, l'*Orchestrator* coordina tanto la pipeline composta da sistemi di estrazione quanto il catalogo contenente i metadati estratti.

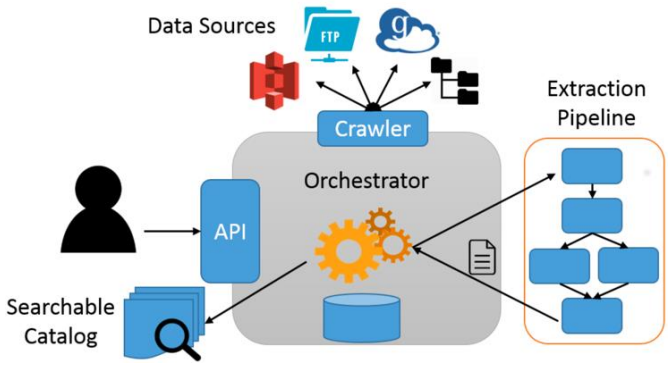

Figura 7. La figura mostra l'architettura di Skluma. (SKLUZACEK, Yler J., KUMAR, Rohan, CHARD, Ryan, HARRISON, Galen, BECKMAN, Paul, CHARD, Kyle, FOSTER, Ian T, "Skluma: An extensible metadata extraction pipeline for disorganized data", in IEEE 14th International Conference on e-Science, 2018;)

In aggiunta, il framework utilizza tecniche di machine learning per determinare dinamicamente i tipi di file trattati, fornisce istruzioni sulla priorità dei file ed esegue una suite dei metadati estratti. Infine, i metadati sono rappresentati nel formato JSON e descrivono il contenuto informativo di ciascun file che può essere successivamente utilizzato per il rilevamento o l'organizzazione. Uno dei vantaggi di Skluma è permettere agli utenti di estendere i tool grazie alla caratteristica modulare dell'architettura e arricchire il complesso di metadati associati alla risorsa[93].

Un ulteriore sistema di estrazione decentralizzata, scalabile e flessibile di metadati che può essere implementato sia in locale che in edge computing è *Xtract*[94]. In generale, *Xtract* è una pipeline di estrazione dinamica e automatica di metadati che coordina l'esecuzione di tali sistemi di estrazione a livello locale o periferico. In particolare, l'organizzazione di analisi da parte della pipeline prima prevede un'applicazione crawler che elabora i dati e definisce la lista dei metadati estratti contenente le proprietà del file nel file system (percorso, dimensione, estensione). Successivamente, invoca un sistema di estrazione di *FileType* su ciascun file analizzato e a partire dal valore ottenuto seleziona dinamicamente sistemi di estrazione aggiuntivi da applicare in base all'output. Nel caso di elaborazione distribuita, gli Endpoint definiscono le risorse di elaborazione e possono essere distribuiti su diversi provider, come *Internet of Things*. In tal senso, gli endpoint consentono di eseguire funzioni di

[93] SKLUZACEK, Yler J., KUMAR, Rohan, CHARD, Ryan, HARRISON, Galen, BECKMAN, Paul, CHARD, Kyle, FOSTER, Ian T, "Skluma: An extensible metadata extraction pipeline for disorganized data", in IEEE 14th International Conference on e-Science, 2018;

[94] SKLUZACEK, Tyler J., "Dredging a Data Lake: Decentralized Metadata Extraction", in Middleware Doctoral Symposium, 2019;

estrazione di metadati in qualsiasi endpoint selezionato[95] consentendo un'analisi distribuita dei dati diminuendone il tempo di elaborazione.

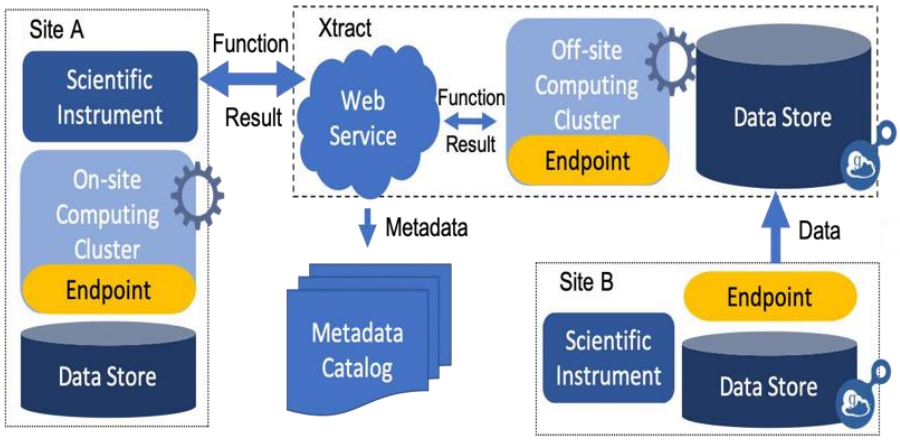

Figura 8. L'architettura della pipeline di Xtract. (SKLUZACEK, Tyler J., "Dredging a Data Lake: Decentralized Metadata Extraction", in Middleware Doctoral Symposium, 2019;

Tuttavia, è necessario sottolineare che *Xtract* è un servizio non ancora implementato ma solo definito dal punto di vista logico e infrastrutturale. In tal senso, i sistemi di estrazione automatica distribuita sembrano fornire una prima soluzione alla gestione dei metadati relativi ai Big Data (ciò ha indotto Greenberg a coniare il termine *Big Metadata*)[96]. Tuttavia, le ricerche e i risultati ottenuti in tale ambito non hanno ancora raggiunto una maturità sufficiente. Per esempio, l'assenza di modelli di metadati in tale dominio determina un importante limite all'interoperabilità semantica mentre l'uso di formati proprietari rischia di trasformare il flusso di bit in un patrimonio conoscitivo illeggibile. Per scongiurare tali rischi sono state sviluppate importanti iniziative dalla comunità di standardizzazione IEEE in termini di gestione dei metadati nel contesto Big Data. Ad esempio, il primo 1st IEEE Big Data Initiative (BDI) Standard Workshop (BDISW) ha previsto la formazione di gruppi di ricerca per sviluppare il "Metadata Standard for Big Data Management"[97] e il "Big Data Metadata Standards"[98]. In particolare, la definizione di modelli di metadati (per esempio il Dublin Core) garantisce la standardizzazione e la valorizzazione dei dati, mentre l'utilizzo di formati aperti (come XML o RDF) garantisce la leggibilità e l'interoperabilità sintattica e semantiche delle risorse.

La maggior parte delle soluzioni di estrazione di metadati sono incentrate su oggetti digitali "statici" in ambienti centralizzati. In particolare, uno tra i prodotti della ricerca maggiormente analizzato sono gli articoli scientifici. Le ragioni di tale scelta dipendono principalmente dalla ricchezza informativa e dalla complessità di analisi di

[95] Ibidem
[96] GREENBERG, Jane, "Big Metadata, Smart Metadata, and Metadata Capital: Toward Greater Synergy Between Data Science and Metadata", Journal of Data and Information Science, 2017;
[97] IEEE, "Standards (IEE BIG DATA)", link: https://bigdata.ieee.org/standards, Consultato il 27/02/2023;
[98] Ibidem

tali oggetti informativi. Un'ulteriore ragione è legata alla scelta dei formati adottati per le pubblicazioni scientifiche, tra cui il *.PDF[99], la cui apertura delle specifiche tecniche permette di conoscere le caratteristiche tecnologiche intrinseche di qualunque documento e svilupparne sistemi di analisi customizzati, aumentando la precisione e la qualità dei risultati. Tuttavia, l'evoluzione tecnologica richiede un'attenzione costante ai sistemi di estrazione automatica di metadati.

Pertanto, definire lo stato dell'arte in merito alle soluzioni attualmente disponibili costituisce il primo fondamentale passaggio per disporre di un quadro conoscitivo complessivamente esaustivo anche al fine di raggiungere l'obiettivo finale del presente lavoro di tesi. Per definire tale quadro si è reso necessario acquisire, filtrare e analizzare una cospicua quantità di letteratura scientifica[100] selezionata da riviste scientifiche. In particolare, la selezione si è basata su misure standardizzate della qualità come, per esempio, l'*H Index* degli autori e le keywords indicate nelle sezioni dedicate all'interno degli articoli, anche al fine di normalizzare i termini in tale contesto e ridurre il rumore e il silenzio informativo nel tentativo di ottenere una più precisa corrispondenza dei risultati. L'analisi ha rivelato lo sviluppo totale di 60 sistemi di estrazione automatica di metadati. Per ciascun tool è stata verificata la reperibilità, la possibilità di utilizzo, la funzione, i formati impiegati e la qualità (intesa come adeguata corrispondenza semantica tra il metadato e il relativo valore) delle informazioni estratte. In particolare, i risultati (illustrati in figura 1) mostrano che il 72% dei sistemi sono "Reperibili" mentre il 28% sono "Irreperibili".

[99] ISO 32000-1:2008 Document management — Portable document format — Part 1: PDF 1.7

[100] Un utile riferimento di partenza è: PARK Jung-ran, BRENZA, Andrew, "Evaluation of Semi---Automatic Metadata Generation Tools: A Survey of The Current State of The Art*",* in *Information technology and libraries*, September 2015. DOI: 10.6017/ital.v34i3.5889

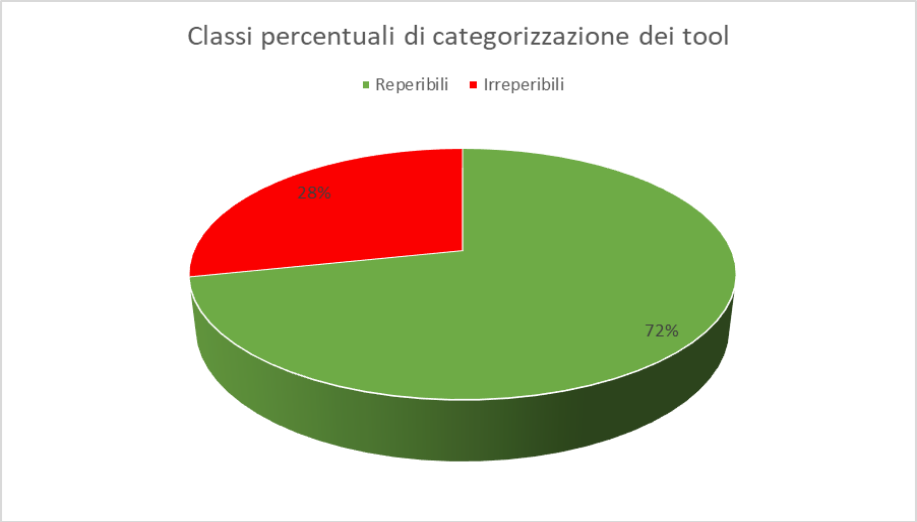

Figura 9. Il grafico illustra la percentuale di tool "Reperibili" e "Irreperibili"

La prima classe indica i tool che sono evidentemente disponibili anche ai fini del loro utilizzo, la seconda classe include i tool che risultano inaccessibili e quindi inutilizzabili. Tra le principali ragioni della irreperibilità figurano l'inesistenza del sito o la migrazione su un nuovo sito attualmente inaccessibile.
È illustrato di seguito a titolo esemplificativo il risultato del censimento dei vari tool:

Type	Name	Location	Techniques
tool	ANVL/ERC Kernel Metadata Conversion Toolkit	http://search.cpan.org/~jak/File-ANVL/anvl	meta-tag harvester
tool	Apache POI - Text Extractor	http://poi.apache.org/download.html	content extractor; metatag harvester; extrinsic auto-generator
tool	Apache Tika	http://tika.apache.org/	content extractor; metatag harvester; extrinsic auto-generator
tool	Ariadne Harvester	http://sourceforge.net/project/ariadnekps/files/?sourc	meta-tag harvester
tool	BIBFRAME Tools	http://www.loc.gov/bibframe/implementation	meta-tag harvester
tool	Data Fountains	http://datafountains.ucr.edu/	content extractor;automatic indexer; metatag harvester; extrir
tool	Dublin Core Meta Toolkit	http://sourceforge.net/project/dcmetatoolkit/files/?sou	meta-tag harvester
tool	Dspace	http://www.dspace.org/	meta-tag harvester; extrinsic auto-generator; social tagging
tool	Editor-Convertes Dublin Core Metadata	http://www.library.kr.ua/dc/dceditunie.html	meta-tag harvester; extrinsic auto-generator; content extractor
tool	Embedded Metadata Extraction Tool (EMET)	http://www.artstor.org/global/g-html/download-emet-	content extractor; metatag harvester; extrinsic auto-generator
tool	Firefox Dublin Core Viewer Extension	http://www.splintered.co.uk/experiments/73/	meta-tag harvester; extrinsic auto-generator;
tool	MarcEdit	http://www.marcedit.reeset.net/	meta-tag harvester
tool	Metatag Extractor Software	http://meta-tag-extractor.software.informer.com/	meta-tag harvester
tool	My Meta Maker	http://www.old.isn-oldenburg.de/services/mmm/	meta-tag harvester
tool	PhotoRDF-Gen	http://www.webposible.com/utilidades/photo_rdf_gen	meta-tag harvester
tool	PyMarc	http://github.com/edsu/pymarc	meta-tag harvester
tool	RepoMMan	http://www.hull.ac.uk/esing/repomman/index.html	content extractor; metatag harvester; extrinsic auto-generator
tool	SHERPA/RoMEO	http://www.sherpa.ac.uk/romeo/api.html	meta-tag harvester
tool	URL and Metatag Extractor	http://www.metatagextractor.com/	meta-tag harvester
tool	Apache Standol	http://www.stanbol.apache.org/	content extractor; automatic indexer

Figura 10. La tabella illustra a titolo esemplificativo alcuni dei tool classificati ciascuno con il colore di appartenenza secondo la reperibilità

Ciascun tool è stato sviluppato secondo specifici obiettivi classificati in "*meta-tag harvester*", "*content extractor*", "*extrinsic auto-generator*", "*automatic indexer*", "*social*

tagging" anche in rapporto al formato di input analizzato. In particolare, la classe "*content extractor*" include anche l'elenco di tool sviluppati per l'estrazione di metadati da diversi oggetti informativi. Infatti, la complessità informativa che contraddistingue i prodotti della ricerca dipende principalmente dall'eterogeneità delle forme e dei contenuti che possono figurare, oltre al testo, come, per esempio le tabelle, i grafici e le immagini. Per migliorare l'esperienza dell'estrazione di metadati sono stati implementati tool e framework anche per l'estrazione da tali oggetti informativi. Tuttavia, i risultati ottenuti da framework in grado di estrarre metadati da più oggetti sono meno significativi rispetto agli strumenti implementati per estrarre metadati esclusivamente da un singolo oggetto. L'analisi dei framework di estrazione automatica di metadati incentrata sui parametri della reperibilità, della funzione, dei formati, e della qualità delle informazioni estratte mostra che tra i sistemi internazionali maggiormente performanti figurano *CERMINE*, sviluppato dall'Università di Varsaia, *ParsCit*, sviluppato dall'Università di Singapore, *FITS*, implementato dall'Università di Harvard e *GROBID* che estrae metadati in formato XML-TEI. Ciascuno dei framework è implementato su specifiche metodologie e tecnologie[101]. Di seguito, si fornisce una panoramica sull'architettura, i formati di input e output e infine i vantaggi e gli svantaggi di ciascun sistema.

1.4.1. CERMINE

«*CERMINE* [Content Extractor and MINEr] *is a comprehensive open-source system for extracting structured metadata from scientific articles in a born-digital form*»[102]. L'architettura del sistema (si veda figura 5) è incentrata su un workflow modulare implementato su un approccio bottom-up con cui sono estratti in ordine i caratteri, le parole, le linee, le zone, le pagine e le coordinate della rappresentazione gerarchico-geometrica del file. I risultati estratti costituiscono l'input dell'analisi successiva che classifica le zone secondo quattro categorie generali: *metadata*, *references*, *body* or *other*. Dalla zona classificata con l'etichetta "*metadata*" è estratto un ricco set di metadati descrittivi rispetto ad altri sistemi di estrazione (come *GROBID* e *ParsCit*), mentre dalla zona classificata con l'etichetta "*references"* sono estratti metadati specificatamente bibliografici. In particolare, CERMINE è in grado di estrarre il *Tittle, Author, Affiliation, Affiliation's metadata, Author–affiliation, E-mail address, Author e-mail, Abstract, Keywords, Journal, Volume, Issue, Pages range, Year, DOI, Reference* e *Reference's metadata*. Dal punto di vista implementativo, gli algoritmi di intelligenza

[101] Le metodologie e tecnologie adottate dai sistemi di estrazione automatica di metadati sono illustrati dettagliatamente nel par. 2.1. del presente lavoro

[102] TKACZYK, Dominik, SZOSTEK, Paweł, FEDORYSZAK, Mateusz, DENDEK, Piotr Jan, BOLIKOWSKI, Łukasz, "CERMINE: automatic extraction of structured metadata from scientific literature", in *International Journal on Document Analysis and Recognition*, Springer, 2015

artificiale[103] che compongono l'architettura di CERMINE spaziano principalmente da algoritmi sviluppati su regole euristiche (bottom-up heuristic-based) ad algoritmi di machine learning (Support Vector Machine) fino ad algoritmi dedicati al data-mining (K-means clustering). Il sistema accetta in input il formato PDF e restituisce i metadati nel formato XML strutturati secondo le regole sintattiche dello standard JATS (Journal Article Tag Suite). La scelta di elaborare in input file in PDF è determinata dalla diffusa applicazione di tale formato, mentre l'XML assicura una maggiore leggibilità nel tempo delle risorse grazie alla sua indipendenza da tecnologie hardware e software, l'interoperabilità tecnologica grazie all'apertura delle specifiche tecniche e quella semantica grazie alla strutturazione sintattica secondo il formato JATS.

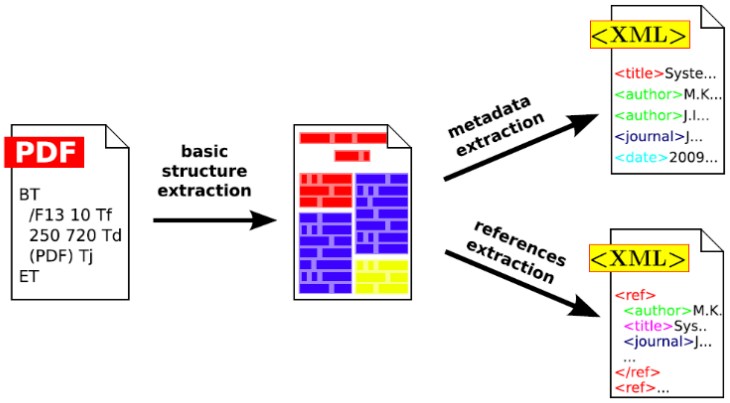

Figura 11. L'architettura del workflow di estrazione automatica di metadati su cui è incentrato CERMINE. TKACZYK, Dominik, SZOSTEK, Paweł, FEDORYSZAK, Mateusz, DENDEK, Piotr Jan, BOLIKOWSKI, Łukasz, "CERMINE: automatic extraction of structured metadata from scientific literature", in *International Journal on Document Analysis and Recognition*, Springer, 2015

Inoltre, è liberamente accessibile sia in versione *web application* che in *standalone application*. Tuttavia, l'applicazione di CERMINE durante il percorso di ricerca ha permesso di dimostrare almeno tre limiti principali:

- La qualità dei metadati estratti da articoli scientifici con layout differenti è minore rispetto a quelli ottenuti dall'analisi del campione per la fase di training incentrato su 125.000 documenti reperiti dall'archivio di PubMed Central;
- L'estrazione delle parole chiavi (*keywords*) è basata sui termini presenti nella sezione "*Keywords*" dell'articolo scientifico. Se tale sezione si trova in una zona differente del testo o è assente, il sistema non è in grado di estrarre tali termini con ricadute negative ai fini dell'accesso. Tale logica è applicabile ad ogni metadato ma assume un'importanza rilevante in considerazione del ruolo ricoperto dalle *keywords* ai fini della ricerca;

[103] Per maggiori dettagli in merito agli algoritmi di intelligenza artificiale, alle metodologie e alle tecniche impiegati nell'estrazione automatica di metadati si veda il par. 2.1. del presente lavoro

- L’estrazione automatica di metadati si concentra sul contenuto informativo testuale e ignora gli ulteriori oggetti che compongono il documento (es. le tabelle, i grafici o le immagini).

La dipendenza di CERMINE, e in generale dei sistemi di estrazione automatica di metadati, dal layout dei documenti costituisce uno degli argomenti ampliamente affrontato nel par. 2.1. del presente lavoro. In questa sede, ci limitiamo a evidenziare che sarebbe necessario costruire un campione rappresentativo dei diversi layout sebbene ciò non risolva totalmente tale problematica. Il secondo e il terzo punto richiederebbero l’estensione con soluzioni implementate allo scopo di estrarre conoscenza dagli oggetti informativi ignorati. Per esempio, nel secondo caso si potrebbe prevedere l’uso di algoritmi di NLP che estraggano le keywords direttamente dal testo indipendentemente da ciò che è indicato nella sezione dedicata. Ciò permetterebbe anche di confrontare la consistenza del contenuto informativo con i termini indicati. Inoltre, se i termini estratti fossero legati ad un vocabolario controllato del dominio di appartenenza si affinerebbe la qualità dell’*information retrieval* diminuendo il silenzio o il rumore informativo. Come nel secondo caso, anche l’estrazione da ulteriori oggetti informativi potrebbe essere ottimizzata estendendo il workflow modulare a strumenti sviluppati con tale finalità.

1.4.2. FITS

FITS (File Information Tool Set) è un framework di estrazione automatica di metadati sviluppato dall’Università di Harvard che identifica, convalida ed estrae i metadati tecnici da diversi formati di file[104]. FITS è implementato in Java (compatibile con Java 1.8 o versioni successive) e fornisce l’output in formato XML.

L’architettura si compone di dodici tool ciascuno con una finalità specifica di seguito illustrata:

[104] Università di Harvard, “File Information Tool Set (FITS). Documentation and official code releases of the FITS and the FITS Web Service projects (File Information Tool Set (FITS). Documentation and official code releases of the FITS and the FITS Web Service projects”, link: https://projects.iq.harvard.edu/fits/home . Consultato il 27/02/2023;

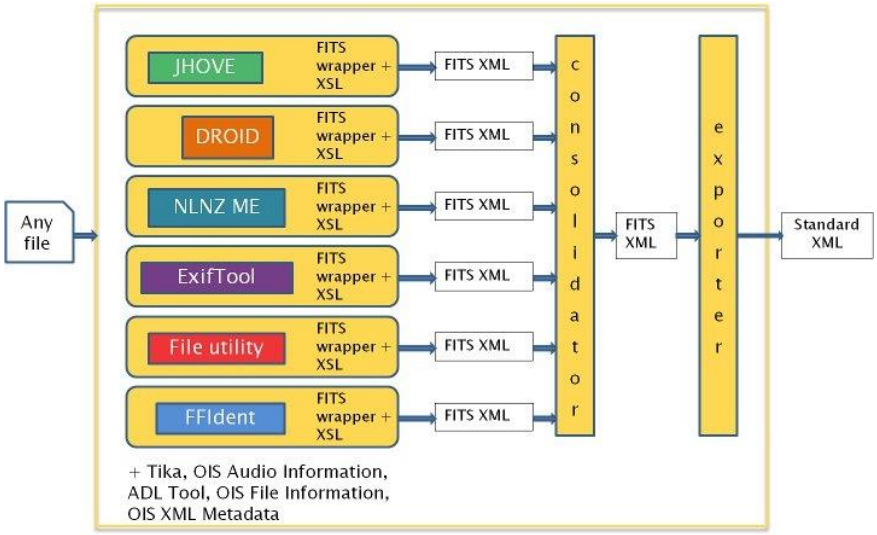

Figura 12. La figura mostra l'architettura generale del sistema FITS (Università di Harvard, "File Information Tool Set (FITS). Documentation and official code releases of the FITS and the FITS Web Service projects (File Information Tool Set (FITS). Documentation and official code releases of the FITS and the FITS Web Service projects", link: https://projects.iq.harvard.edu/fits/home . Consultato il 27/02/2023)

In particolare, i tool che compongo il FITS sono:

- **ADL Tool**, è un tool di estrazione automatica di metadati da file Audio Decision List sviluppato dalla Biblioteca di Harvard;
- **Apache Tika**, è un tool che identifica diversi formati di file (tra cui HTML, XML, OOXML, ODF, PDF, ZIP, TXT, MP3, MP4, FLV, JPEG, TIFF, MBOX) sviluppato e mantenuto da Apache;
- **Droide**, è un tool scritto in Java che identifica i formati di file sviluppato dai UK National Archives;
- **ExifTool**, è un tool scritto in Perl che identifica ed estrae i metadati tecnici sviluppato da Phil Harvey;
- **Ffident**, è un tool scritto in Java che identifica i formati di file;
- **File utility**, è un tool in bundle con Linux, UNIX e OS X. GnuWin32 che identifica i file;
- **Jhove**, è un tool scritto in Java sottoposto a controllo dalla Open Preservation Foundation che identifica, estrae metadati tecnici e convalida i file;
- **MediaInfo**, è un tool scritto in C++ eseguibile anche tramite Java dalla MediaArea.net che identifica ed estrae i metadati tecnici per i file video;
- **National Library of New Zealand (NLNZ) Metadata Extractor Tool**, è un tool scritto in Java mantenuto dalla National Library of New Zeland che identifica ed estrae metadati tecnici da diversi formati, tra cui PG, TIFF, GIF, BMP, WAV, MP3, XML, HTML, PDF, DOC, WORDPERFECT, MSWORKS, ODT;

- **OIS Audio Information,** è un tool che estrae metadati tecnici da formati di file audio sviluppato dalla Library of Harvard;
- **OIS File Information**, è un tool che estrae metadati tecnici dai file sviluppato dalla Library of Harvard;
- **OIS XML Information**, è un tool sviluppato dalla Library of Harvard che identifica ed estrae i metadati tecnici da testo in formato FITS.

I metadati tecnici estratti da FITS offrono all'utente un complesso di informazioni essenziali per la conservazione digitale. In particolare, l'identificazione e l'estrazione delle informazioni in merito al formato permettono di supportare la riproducibilità e leggibilità delle risorse nel tempo e ciò fornisce un utile supporto alla decisione in merito all'eventuale riversamento. Inoltre, l'estrazione di metadati come il *checksum* contribuisce alla verifica dell'integrità del contenuto informativo e, infine, la caratteristiche modulare permette di estendere l'architettura con l'aggiunta di ulteriori strumenti arricchendo ulteriormente il set di metadati.

L'implementazione di sistemi di estrazione automatica di metadati costituisce un utile supporto ai ricercatori poiché permette la semplificazione della descrizione dei prodotti della ricerca e agli istituti di ricerca perché migliora l'interoperabilità tra repository attraverso l'impiego di standard descrittivi (per esempio JATS) e formati aperti (per esempio XML). Inoltre, l'apertura delle specifiche tecniche e l'indipendenza dalle architetture hardware e software ne permettono la libera adozione anche nel tempo, mentre l'utilizzo di algoritmi di intelligenza artificiale ottimizza il processo di valorizzazione dei campi, che se compilati manualmente, risulterebbe essere un'attività altamente onerosa e *time-consuming*. Ciò che ne consegue è la valorizzazione delle risorse grazie al miglioramento della qualità dell'information retrieval e all'aumento della disseminazione scientifica con ricadute positive anche in merito alla gestione dei fondi pubblici. Tuttavia, il contesto non esula anche da aspetti svantaggiosi. L'estrazione della struttura dipende principalmente dal layout impiegato, per cui lo stesso contenuto ma espresso con layout differenti potrebbe non essere estratto oppure assegnato al metadato non corretto. In tal senso, la dipendenza dal singolo layout di documenti ne determina l'alta o la scarsa qualità delle informazioni anche in relazione alla eventuale coerenza con uno o più domini.

Sulla base delle considerazioni fin qui effettuate si rende necessario affinare ed estendere gli strumenti di estrazione per ottenere un set di metadati adeguatamente valorizzati. In particolare, sarebbe opportuno addestrare gli algoritmi di intelligenza artificiale all'analisi di diversi modelli per la rappresentazione delle informazioni. Pertanto, tale addestramento dovrebbe prevedere la costruzioni di campioni composti da documenti basati su diversi layout. L'aggiunta di strumenti per l'estrazione di metadati da ulteriori oggetti oltre il testo che potrebbero arricchire il set di metadati, mentre l'inclusione di tool per l'estrazione di metadati tecnici (per esempio FITS) e l'uso

di formati aperti come XML supporterebbero il processo di conservazione digitale nel tempo delle risorse.

Altri strumenti permettono di estrarre automaticamente metadati e di rappresentarli direttamente in una struttura di tipo XML. Tra questi giova ricordare GROBID che è una libreria sviluppata nel 2008 in Java «*for extracting, parsing and re-structuring raw documents such as PDF into structured XML/TEI encoded documents with a particular focus on technical and scientific publications*»[105]. L'obiettivo finale del framework consiste nell'estrazione e nella trasformazione dei dati non-strutturati in formato strutturati.

Tale trasformazione determina le ricadute espresse nell'illustrazione seguente che include anche la rappresentazione generica dell'architettura di GOBRID:

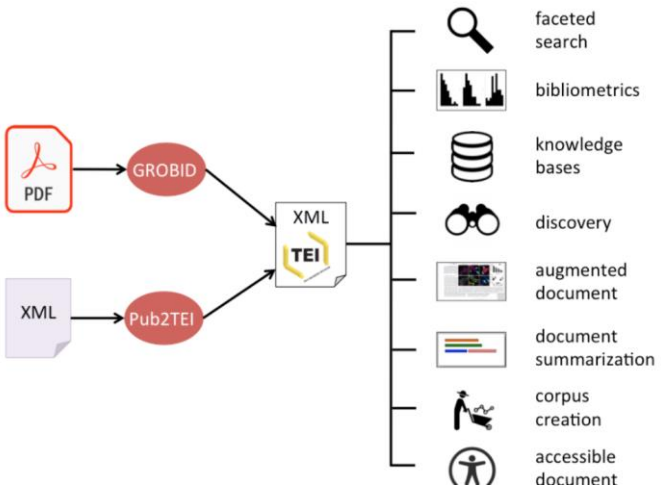

*Figura 13. Rappresentazione generale dell'architettura di GROBID e delle relative ricadute (*LOPEZ, Patrice, "GROBID", link: https://grobid.readthedocs.io/en/latest/Introduction/ Consultato il 27/02/2023*)*

La recente adozione da parte degli editori di adottare il formato XML ha portato gli sviluppatori di GROBID all'estensione dell'architettura grazie all'utilizzo del tool *Pub2TEI* per l'elaborazione del formato XML in XML/TEI. La metodologia utilizzata per l'estrazione di metadati da GROBID è incentrata sull'analisi sul *document parsing* per la generazione di modelli finalizzati alla classificazione delle sequenze di testo[106] e all'estrazione dei metadati. Gli algoritmi di machine e deep learning permettono di affinare la qualità dei risultati estratti valorizzando il contenuto informativo in termini di recupero e ascesso alle risorse. Tuttavia, come nel caso di CERMINE, tale qualità è limitata dalla dipendenza del layout utilizzato da ciascun documento, mentre l'estrazione da layout differenti determina una scarsa capacità di estrazione o una insufficiente qualità a causa della dipendenza dalla rappresentazione formale.

[105] LOPEZ, Patrice, "GROBID", link: https://grobid.readthedocs.io/en/latest/Introduction/ Consultato il 27/02/2023;
[106] La classificazione è realizzata mediante algoritmi di Machine Learning (tra cui CRF visti in ParsCit) e Deep Learning.

Interessante risulta, inoltre, l'approccio di alcuni tool che permettono di lavorare sull'estrazione automatica di metadati da stringhe di reference. Tali strumenti potenziano la possibilità di rappresentare elementi informativi interni ed esterni ai documenti in chiave di relazione. Ad esempio, ParsCit è un tool open-source sviluppato nel 2008 dalla National University of Singapore (NUS) in collaborazione con la Penn State University su diversi linguaggi, tra cui Perl, HTML, CSS, XSLT, Ruby, Raku e altri[107]. Obiettivo finale di ParsCit è «*recovering bibliographic and document structure metadata from scholarly documents*»[108] attraverso algoritmi di intelligenza artificiale incentrati sulla tecnica dei *conditional random fields* (CRF) da documenti nativi digitali in formato PDF. In particolare, l'estrazione riguarda le informazioni in merito all'autore, alla data, al titolo, all'abstract, all'introduzione e alla metodologia. Dal punto di vista metodologico, l'High Order Semi-CRF (HO-SCRDFs) prima modella la probabilità di sequenza di etichette di token finiti attraverso il modulo *Linear chain CRF* (L-CRF), successivamente il modulo *Semi-CRFs* estendo tale modellazione a una lunghezza variabile. Le prestazioni sono convalidate attraverso tre attività di estrazione di informazioni da documenti accademici:

- **Reference String Parsing**, tokenizza ed etichetta i singoli token che compongono le stringe di *References*;
- **Generic Section Labeling**, recupera la struttura logica delle sezioni principali di testo accademico estraendo le etichette;
- **Author and Affiliation Extraction**, estrae le occorrenze dell'autore dalla sezione dell'intestazione del documento.

In tal senso, l'estrazione di informazioni dalle *references* da risorse accademiche permette ai ricercatori di acquisire velocemente la letteratura scientifica per poter definire lo stato dell'arte nel contesto di ricerca. Tuttavia, è necessario che tali informazioni siano accompagnate da ulteriori informazioni utili a descrivere il contenuto del documento al fine di garantire la conservazione digitale e l'accesso nel tempo.

[107] COUNCILL, Isaac, GILES, Lee C., KAN, Min-Yen, "ParsCit", link: https://github.com/knmnyn/ParsCit. Consultato il 27/02/2023

[108] CUONG, Nguyen Viet, CHANDRASEKARAN, Muthu Kumar, KAN, Min-Yen, LEE, Wee Sun, "Scholarly Document Information Extraction using Extensible Features for Efficient Higher Order Semi-CRFs", in JCDL'15, ACM, 2015. DOI: http://dx.doi.org/10.1145/2756406.2756946

2. La costruzione del framework per l'estrazione automatica dei metadati dai prodotti della ricerca

2.1. Modelli e tecnologie per la definizione del framework di estrazione automatica di metadati

2.1.1. Il modello italiano per la conservazione degli oggetti e dei documenti digitali

La conservazione digitale è un processo necessario della corretta dematerializzazione dei flussi documentali per garantire l'autenticità, l'affidabilità, l'integrità e la possibilità di utilizzo degli oggetti digitali nel tempo. In particolare, il modello italiano per la formazione degli archivi digitali si fonda su un complesso quadro normativo incentrato sul D. Lgs. 7 marzo 2005 n. 82 recante il Codice dell'Amministrazione Digitale (CAD) e sulle attuative Linee Guida sulla formazione, gestione e conservazione dei documenti informatici emanate da AgID attuabili dalle pubbliche amministrazioni e dai soggetti privati. Tali disposizioni dovrebbero armonizzare le metodologie e le tecnologiche adoperate per la conservazione degli oggetti digitali. Tuttavia, la relativa attuazione impatta sulle difficoltà intrinseche determinate dall'evoluzione tecnologica, tra cui l'obsolescenza dei formati e la gestione dei documenti come unità documentarie svincolate dagli altri oggetti digitali logicamente legati. Ciò da un lato determina il rischio della illeggibilità del patrimonio archivistico prodotto su supporti informatici e dall'altro la mancata definizione del vincolo archivistico su cui si fonda il concetto stesso di archivio. Accanto a tali problematicità si aggiunge la gestione della complessa questione legata alla creazione di archivi ibridi che implica la scissione fisica dei documenti prodotti su supporto cartaceo da quelli informatici ma entrambi afferenti al medesimo archivio o addirittura alla medesima pratica. Un ulteriore norma volta a tutelare gli archivi dichiarati di notevole valore storico prodotti dalle pubbliche amministrazioni e dai soggetti privati è il D. Lgs. 22 gennaio 2004 n. 42 recante il Codice dei beni culturali e del paesaggio. Ciò genera maggiore confusione anche nei non addetti ai lavori nella distinzione delle procedure che fanno capo al Ministero della Cultura o a quelle affidate all'Agenzia per l'Italia Digitale. Come sottolineato da Pigliapoco ciò determina una «differenziazione che si contrappone alla concezione dell'archivio come complesso unitario, organico, dei documenti redatti o ricevuti da una persona fisica o giuridica nel corso delle sue attività indipendente dal supporto sul quale sono prodotti e genera una condizione di incertezza sia nei soggetti produttori, i quali finiscono per trattare i documenti informatici come oggetti a sé stanti garantendo la conservazione "a norma" solo di quelli aventi rilevanza giuridica o per i quali esiste un obbligo normativo esplicito»[109].

Inoltre, alle disposizioni nazionali si aggiungo i regolamenti emanati dall'Unione Europea, tra cui il Regolamento eIDAS (electronic IDentification Authentication and Signature) n° 910/2014 sull'identità digitale che ha l'obiettivo di fornire una base

[109] PIGLIAPOCO, Stefano, "La Conservazione Digitale in Italia. Riflessioni su modelli, criteri e soluzioni", in *JLis.it,* 2019. DOI: 10.4403/jlis.it-12521;

normativa a livello comunitario per i servizi fiduciari e i mezzi di identificazione elettronica degli stati membri e il Regolamento GDPR (General Data Protection Regulation) n° 679/2016 relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati. Infine, alle disposizioni generali si aggiungo al modello italiano della conservazione digitale quelle della conservazione di specifiche tipologie documentali (si pensi per esempio al DPR 11 febbraio 2005, n.68 recante le disposizioni per l'utilizzo e la conservazione della posta elettronica certificata) e le relative complicazioni che ne conseguono[110].

La complessità del quadro normativo che disciplina ogni aspetto del ciclo di vita delle risorse lascia dedurre che il modello italiano per la conservazione sia definitivamente delineato. Tuttavia, la traduzione delle norme nel contesto applicativo impatta con i molteplici limiti metodologici e tecnologici adottati dalle realtà interessate.

Definire il quadro normativo di riferimento della conservazione digitale per evidenziare le principali difficoltà nella relativa attuazione e avanzare eventuali proposte al fine di contribuire a migliorare l'applicazione delle disposizioni normative, anche attraverso il supporto di soluzioni tecnologiche, è il fine del presente paragrafo.

L'art. 30 del stabilisce che «lo Stato, le regioni, gli altri enti pubblici territoriali nonché' ogni altro ente ed istituto pubblico hanno l'obbligo di garantire la sicurezza e la conservazione dei beni culturali di loro appartenenza».[111] In particolare, l'art. 2 comma 2 del medesimo codice chiarisce il significato di bene culturale che include «le cose immobili e mobili che, ai sensi degli articoli 10 e 11, presentano interesse artistico, storico, archeologico, etnoantropologico, archivistico e bibliografico e le altre cose individuate dalla legge o in base alla legge quali testimonianze aventi valore di civiltà»[112]. Ciò obbliga le pubbliche amministrazioni alla conservazione dei documenti prodotti nel pieno svolgimento delle proprie funzioni giuridiche anche quelli formati su supporto informatico le cui modalità di conservazione sono disciplinate dal D. Lgs. 7 marzo 2005 n. 82 recante il Codice dell'Amministrazione Digitale (CAD). L'art. 43 comma 1-bis del CAD stabilisce infatti che «gli obblighi di conservazione e di esibizione di documenti si intendono soddisfatti a tutti gli effetti di legge a mezzo di documenti informatici, se le relative procedure sono effettuate in modo tale da garantire la conformità ai documenti originali e sono conformi alle Linee guida»[113]. In tal senso, l'art. 71 dello stesso Codice sancisce che «l'AgID, previa consultazione pubblica da svolgersi entro il termine di trenta giorni, sentiti le amministrazioni competenti e il Garante per la protezione dei dati personali nelle materie di competenza, nonché acquisito il parere della Conferenza unificata, adotta Linee guida contenenti le regole

[110] ROVELLA, Anna, *"*La posta elettronica negli archivi di persona: conservazione e accesso*", in AIDA Informazioni*, 2022: 113-128. DOI: 10.57574/596516316

[111] Parlamento italiano, "D. lgs. recante il Codice dei beni culturali e del paesaggio, ai sensi dell'articolo 10 della legge 6 luglio 2002, n. 137", 22 gennaio 2004, n. 42. (in G. U. n. 45 del 24 febbraio 2004 – Suppl. Ordinario n. 28)

[112] Ibidem

[113] Art. 43 del D. lgs. 7 marzo 2005 n.82 recante il Codice dell'Amministrazione Digitale, https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legislativo:2005-03-07;82 . Consultato il 27/02/2023

tecniche e di indirizzo per l'attuazione del presente Codice»[114]. In virtù di quanto sopra indicato e alla luce della determinazione n. 160/2018 contenente il "Regolamento per l'adozione di linee guida per l'attuazione del Codice dell'Amministrazione Digitale", AgID ha adottato con determinazione n. 407/2020 le nuove "*Linee guida per la formazione, gestione e conservazione dei documenti informatici*"[115] modificate a maggio 2021 con la successiva determinazione n. 371/2021 ed entrate in vigore a partire dal 1° gennaio 2022. In particolare, l'articolo 34, comma 1-bis, prevede che la conservazione dei documenti informatici possa avvenire:

a) all'interno della propria struttura organizzativa;
b) affidandola, in modo totale o parziale, nel rispetto della disciplina vigente, ad altri soggetti, pubblici o privati che possiedono i requisiti di qualità, di sicurezza e organizzazione individuati, nel rispetto della disciplina europea, nelle Linee guida di cui all'art 71 relative alla formazione, gestione e conservazione dei documenti informatici nonché in un regolamento sui criteri per la fornitura dei servizi di conservazione dei documenti informatici emanato da AgID, avuto riguardo all'esigenza di assicurare la conformità dei documenti conservati agli originali nonché la qualità e la sicurezza del sistema di conservazione.".

In particolare, la conservazione dei documenti informatici prodotti dalle pubbliche amministrazioni e affidata a soggetti esterni deve essere adeguatamente configurata nel rispetto delle disposizioni contenute nel regolamento citato nell'articolo 34 del CAD ed emanato nel dicembre 2021 da AgID. In particolare, il "Regolamento sui criteri per la fornitura dei servizi di conservazione dei documenti informatici" definisce i requisiti di qualità, sicurezza e organizzazione e l'eventuale iscrizione facoltativa al marketplace da parte dei soggetti che intendono erogare servizio di conservazione dei documenti informatici per conto delle pubbliche amministrazioni[116]. Attualmente l'elenco degli iscritti al marketplace di AgID conta 50 organizzazioni pubbliche e private[117] che erogano servizio di conservazione altamente qualificato. Tali organizzazioni sono

[114] Art.71 del D. lgs. 7 marzo 2005 n.82 recante il Codice dell'Amministrazione Digitale, https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legislativo:2005-03-07;82 . Consultato il 27/02/2023

[115] La scelta di adottare Linee Guida è finalizzata ad incorporare i DPCM e le circolari contenenti le disposizioni anche in materia di conservazione al fine di rendere agevole la lettura ai soggetti interessati e la relativa applicazione delle medesime disposizioni.

[116] Nel caso di affidamento del servizio di conservazione dei documenti informatici per conto dei soggetti privati non sussiste il vincolo di affidare la conservazione a soggetti che siano in possesso dei requisiti fissati dall'AgID nel Regolamento sui criteri per la fornitura dei servizi di conservazione dei documenti informatici né tanto meno l'iscrizione al marketplace. Se i documenti prodotti dalle pubbliche amministrazioni costituiscano dei potenziali beni culturali di proprietà dello Stato, la scelta di svincolare i soggetti privati dall'obbligo di affidare la conservazione a soggetti non qualificati da un lato semplifica le procedure di conservazione dall'altro rischia di determinare la perdita del patrimonio informativo del tessuto produttivo del paese e sminuire l'importanza degli archivi prodotti dai soggetti privati.

[117] Agenzia per l'Italia Digitale, "Elenco dei conservatori iscritti al Marketplace dei servizi di conservazione", link: https://conservatoriqualificati.agid.gov.it/?page_id=276. Consultato il 27/02/2023;

formalmente denominati come Conservatori. Dal punto di vista strutturale, il regolamento si compone di tre documenti:

- Il testo del regolamento contenente:
  a) i requisiti di qualità, di sicurezza e organizzazione;
  b) le procedure e la durata di iscrizione al marketplace per i servizi di conservazione e verifica dei requisiti;
  c) le modalità di cancellazione dal marketplace;
  d) la procedure di verifica dei requisiti dei conservatori non iscritti al marketplace;
  e) lo schema per la stesura del piano di cessazione del servizio di conservazione.
- L'allegato A relativo a i requisiti per l'erogazione del servizio di conservazione per conto delle pubbliche amministrazioni che definisce anche i profili professionali richiesti ai Conservatori e le relative funzioni, tra cui il Responsabile del servizio di conservazione e il Responsabile della funzione archivistica di conservazione;
- L'allegato B che stabilisce le modalità per la stesura e la gestione del piano di cessazione del servizio di conservazione sia nel caso di cessazione volontaria che involontaria.

Tra i requisiti richiesti ai fornitori di servizi di conservazione, nel rispetto degli artt. 33 e 34 del CAD, delle disposizioni europee e in conformità alle Linee Guida sulla formazione, gestione e conservazione dei documenti informatici figura l'obbligo di configurare il processo di conservazione cosicché:

a) il sistema di conservazione assicuri, per quanto in esso conservato, caratteristiche di autenticità, integrità, affidabilità, leggibilità, reperibilità;
b) i soggetti, pubblici o privati, che erogano il servizio di conservazione per conto delle pubbliche amministrazioni, offrano idonee garanzie organizzative, tecnologiche e di protezione dei dati personali e assicurino la conformità dei documenti conservati agli originali nonché la qualità e la sicurezza del sistema di conservazione[118].

In particolare, il modello funzionale su cui devono essere implementati i sistemi di conservazione digitale dei documenti informatici è lo standard ISO 14721:2012 OAIS (Open Archival Information System - *Sistema informativo aperto per l'archiviazione*) e di seguito illustrato in figura 1.

[118] Agenzia per l'Italia Digitale, "Regolamento sui criteri per la fornitura dei servizi di conservazione dei documenti informatici", Roma, 2021

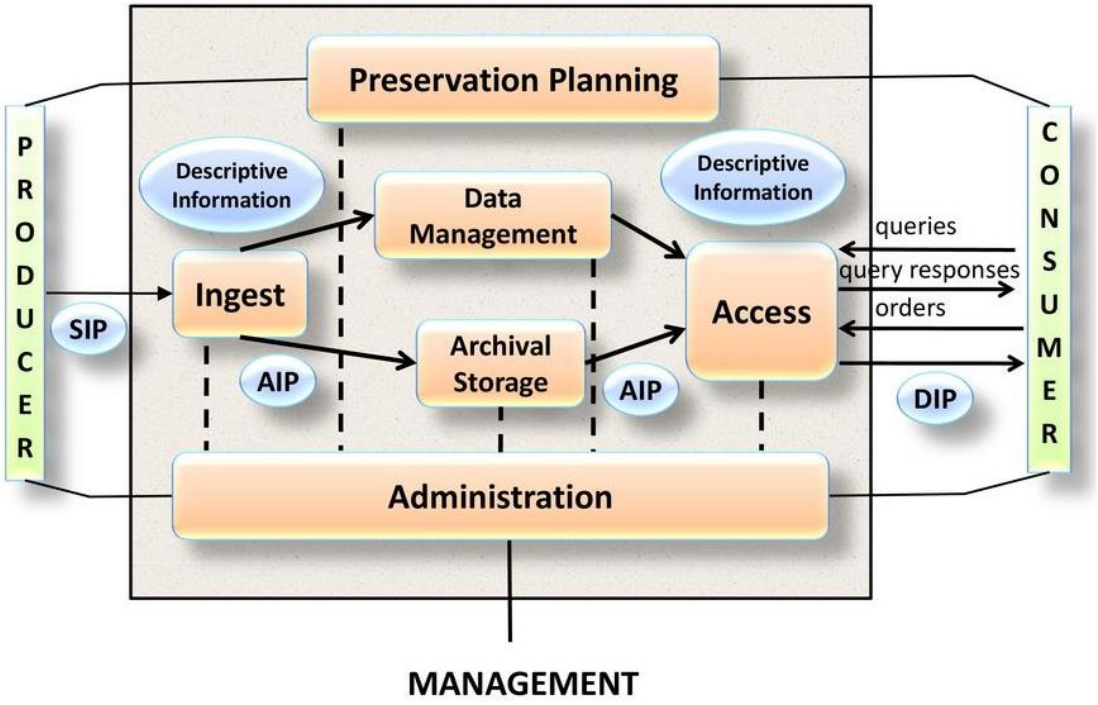

Figura 14. Il modello di riferimento OAIS dei sistemi di conservazione (ISO 14721:2012 OAIS)

La genesi dell'elaborazione dello standard ISO 14721:2012 risale al 1997 da parte del Consultative Committee for Space Data Systems (CCSDS – Comitato consultivo per i sistemi di dati spaziali) denominato *Reference Model for an Open Archival Information System* per la conservazione dei dati digitali a lungo termine generati dalle osservazioni geospaziali. Nel 2003 l'organismo di standardizzazione ISO approva il modello di riferimento quale standard internazionale 14721:2003 OAIS aggiornato successivamente nel 2012. L'ISO 14721:2012 «prevede l'esistenza di un'organizzazione di persone e sistemi, denominata "*archivio OAIS*" che assume la responsabilità della conservazione dei contenuti digitali dei soggetti produttori (*Producer)* e del mantenerla disponibile per una comunità [designata] di utenti (*Consumer*)».[119] Come sottolineato da Giovanni Michetti «il termine aperto si riferisce al fatto che il processo di elaborazione dello standard si è svolto condividendo in ampia misura le conoscenze e utilizzando il contributo di tutti i soggetti interessati»[120] oltre all'apertura delle specifiche tecniche su cui si fonda tale standard. Tali specifiche prevedono che conservazione dei contenuti digitali avvenga mediante la generazione di pacchetti informativi (*Information Packages*) contenenti gli oggetti digitali e le informazioni sulla rappresentazione. I diversi pacchetti informativi assumono una funzione e una

[119] PIGLIAPOCO, Stefano, "La Conservazione Digitale in Italia. Riflessioni su modelli, criteri e soluzioni", in JLis.it, 2019. DOI: 10.4403/jlis.it-12521;
[120] MICHETTI, Giovanni, "Il modello OAIS", DigItalia, Anno III, Numero I, 2008. p.35;

denominazione specifica in rapporto alle varie fasi del processo di conservazione durante le quali sono generati dai soggetti interessati. Tali pacchetti si distinguono in Submission Information Package (SIP), Archival Information Package (AIP) e Dissemination Information Package (DIP). Un ulteriore punto di forza dello standard è l'estrema astrattezza che ne permette l'adesione indipendentemente dal dominio. Tuttavia, ciò richiede un'adeguata contestualizzazione dello stesso standard al fine di garantire la coerenza in fase applicativa. In tal senso, il legislatore italiano ha re-interpretato i ruoli previsti dall'ISO per assicurare una maggiore aderenza alla normativa in vigore in merito alla conservazione digitale. In particolare, paragrafo 4.4. delle Linee AgID individua i soggetti coinvolti nel processo di conservazione (definito nel paragrafo 4.7. delle stesse Linee Guida anch'esso sulla base dello standard ISO) tra cui:

a) Titolare dell'oggetto di conservazione;
b) Produttore dei PdV;
c) Utente Abilitato;
d) Responsabile della Conservazione;
e) Conservatore.

Le definizioni di ciascun ruolo sono fornite dall'Allegato 1 relativo al glossario dei termini e degli acronimi utilizzati nelle Linee Guida. In particolare, il Titolare dell'oggetto di conservazione è definito come il soggetto produttore degli oggetti digitali sottoposti a conservazione mentre il Produttore dei PdV è il soggetto fisico interno al Titolare dell'oggetto di conservazione che produce il pacchetto di versamento ed è responsabile del trasferimento del suo contenuto nel sistema di conservazione. L'utente abilitato è la persona, ente o sistema che interagisce con i servizi di un sistema di gestione informatica dei documenti e/o di un sistema per la conservazione dei documenti informatici, al fine di fruire delle informazioni di interesse, e il responsabile della conservazione è il soggetto che definisce e attua le politiche complessive del sistema di conservazione e ne governa la gestione con piena responsabilità ed autonomia. Tale figura riveste un ruolo centrale nella gestione del processo di conservazione dei documenti informativi del Titolare dell'oggetto di conservazione e rappresenta il principale interlocutore del Responsabile del Servizio di conservazione[121], interno al Conservatore, ovvero il soggetto pubblico o privato che svolge attività di conservazione dei documenti informatici, nel caso di affidamento del servizio di conservazione. Anche i pacchetti informativi previsti dall'OAIS sono stati denominati rispettivamente:

[121] L'Allegato 1 "Glossario dei termini e degli acronimi" alle Linee Guida definisce il Responsabile del servizio di conservazione: "il soggetto che coordina il processo di conservazione all'interno del conservatore, in possesso dei requisiti professionali individuati da AGID [nel Regolamento sui criteri per la fornitura dei servizi di conservazione dei documenti informatici".

- Pacchetto di Versamento (PdV), è il pacchetto informativo inviato dal produttore al sistema di conservazione secondo il formato descritto nel manuale di conservazione;
- Pacchetto di Archiviazione (PdA), è il pacchetto informativo generato dalla trasformazione di uno o più pacchetti di versamento coerentemente con le modalità riportate nel manuale di conservazione;
- Pacchetto di Distribuzione (PdD), è il pacchetto informativo inviato dal sistema di conservazione all'utente in risposta ad una sua richiesta di accesso a oggetti di conservazione[122].

L'interpretazione del legislatore relativamente ai ruoli, alle responsabilità e alla metodologia per la gestione degli oggetti digitali nell'archivio OAIS ha determinato la definizione del modello nazionale della conservazione digitale modellato nelle figure 3 e 4 che rappresentano rispettivamente il modello della conservazione digitale all'interno della struttura organizzativa dell'ente produttore che ricopre anche il ruolo di conservatore e il modello della conservazione digitale affidata all'esterno dell'ente produttore.

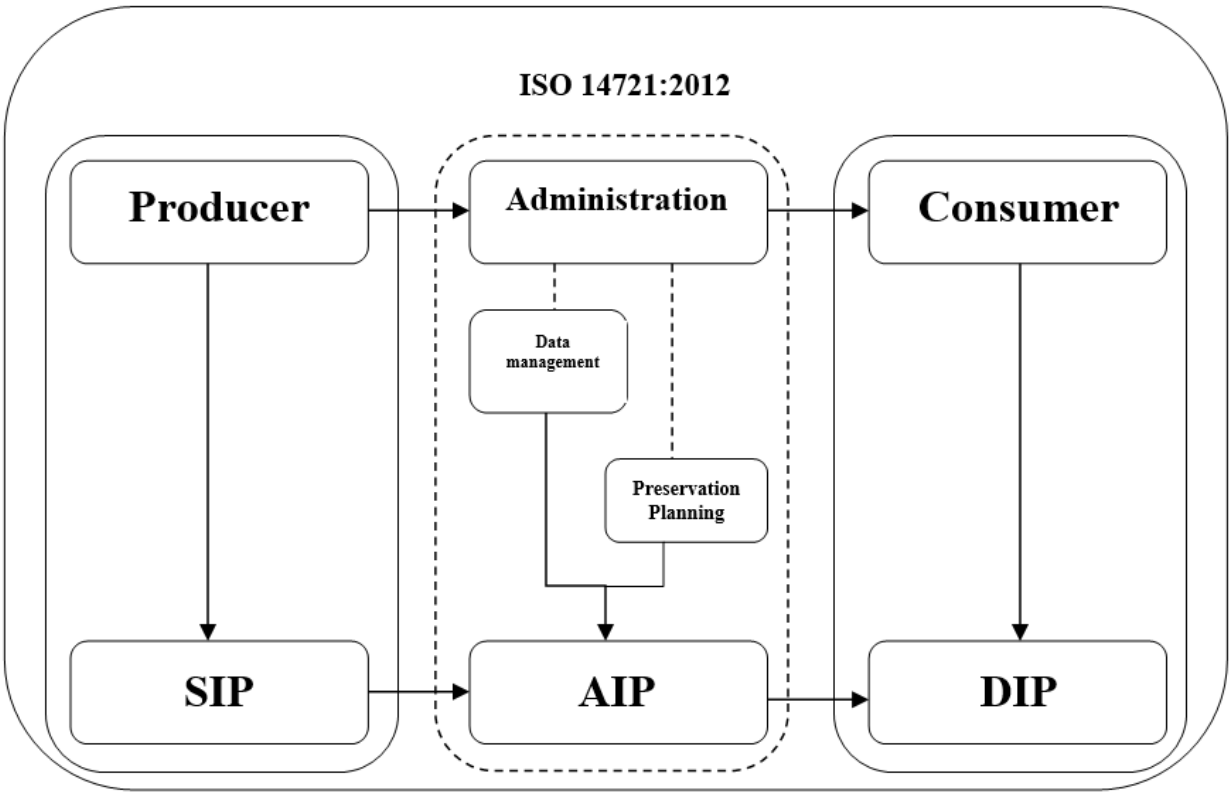

Figura 15. L'archivio OAIS dello standard ISO 14721:2012

[122] Agenzia per l'Italia Digitale, "Allegato 1. Glossario dei Termini e degli Acronimi" in Linee Guida sulla formazione, gestione e conservazione dei documenti informatici, Roma, maggio 2021

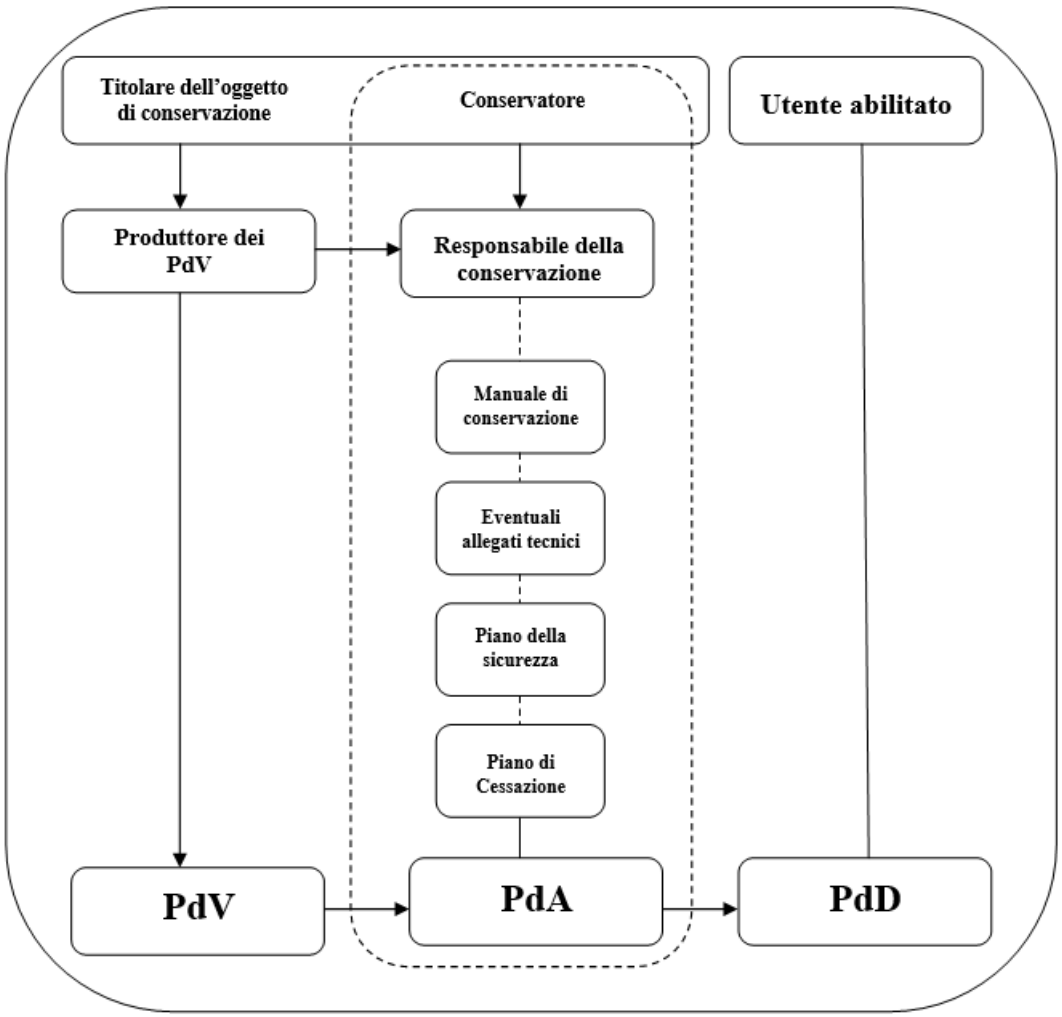

Figura 16. Modello della conservazione digitale in Italia sulla base dello standard ISO 14721:2012 all'interno della propria struttura organizzativa

Il processo di conservazione degli oggetti digitali[123] è definito dal paragrafo 4.7 delle Linee guida secondo il quale la generazione e il trasferimento del pacchetto di versamento, da parte del Produttore dei PdV, al sistema di conservazione del Conservatore designato, secondo le modalità e il formato indicato nel manuale di conservazione[124], e concordato tra le parti, danno avvio a tale processo. Dal punto di

[123] Il glossario dei termini e degli acronimi contenuto nell'Allegato 1 "Glossario dei termini e degli acronimi" delle Linee guida definisce l'oggetto digitale un "*Oggetto informativo digitale, che può assumere varie forme, tra le quali quelle di documento informatico, fascicolo informatico, aggregazione documentale informatica o archivio informatico*"

[124] Il glossario dei termini e degli acronimi contenuto nell'Allegato 1 delle Linee guida definisce il manuale di conservazione il "*documento informatico che descrive il sistema di conservazione e illustra dettagliatamente l'organizzazione, i soggetti coinvolti e i ruoli svolti dagli stessi, il modello di funzionamento, la descrizione del processo, la descrizione delle architetture e delle infrastrutture*". È necessario evidenziare che le nuove Linee Guida hanno esteso indirettamente l'obbligo di redazione del manuale di conservazione anche ai soggetti privati in quanto si rende obbligatoria la nomina formale, anche per questi ultimi, del responsabile della conservazione. In particolare, le responsabilità del responsabile della conservazione possono essere in parte o in tutti delegate all'interno della propria

vista tecnologico il pacchetto di versamento è generalmente trasmesso o attraverso Web Services, o SFTP o Upload manuale e può contenere uno o più oggetti digitali e i relativi metadati[125]. La scelta della tecnologia è concordata tra titolare dell'oggetto di conservazione e conservatore e in tal senso è indispensabile il costante confronto tra il Produttore dei PdV e il Responsabile della conservazione. L'acquisizione del PdV da parte del sistema di conservazione ne permette la verifica relativamente agli oggetto digitali, ai metadati e ai formati scelti tra quelli indicati nell'Allegato 2 delle Linee guida denominato "Formati di file e riversamento". Nel caso in cui il sistema di conservazione dovesse riscontrare delle anomalie sul PdV versato, queste ultime sono istantaneamente comunicate al Produttore dei PdV e al Responsabile della conservazione. Per esempio, tra le soluzioni adoperate per la comunicazione figurano l'interfaccia grafica e l'invio di un messaggio di posta elettronica all'indirizzo designato. In caso di accettazione, il sistema di conservazione prende in carico il PdV e genera, anche in modo automatico, il rapporto di versamento[126] che deve essere identificato in modo univoco dal sistema di conservazione e su cui devono essere calcolate una o più impronte sull'intero contenuto del pacchetto di versamento al fine di garantire l'integrità del contenuto. Inoltre, il rapporto di versamento deve definire i limiti temporali entro cui si colloca il PdV attraverso il riferimento al Tempo universale coordinato (UTC) (per esempio la data di versamento e la data di presa in carico). Tale rapporto di versamento deve essere sottoscritto o dal responsabile della conservazione o dal responsabile del servizio di conservazione e deve essere consultabile dal Produttore dei PdV al fine di verificare che i documenti versati siano stati correttamente accettati dal sistema di conservazione. Il PdV preso in carico dal sistema di conservazione è trasformato in PdA (che può contenere uno o più PdV) e deve essere sottoscritto o dal responsabile della conservazione o dal responsabile del servizio di conservazione oppure sigillato elettronicamente dal Conservatore secondo le modalità previste dalla let. f) del par. 4.7 delle Linee Guida. La gestione del PdA da parte del Conservatore deve essere incentrata sullo standard nazionale UNI 11386:2020 *Supporto all'Interoperabilità nella Conservazione e nel Recupero degli Oggetti digitali* (SInCRO)[127] dettagliato nel paragrafo 1.1 del presente lavoro di tesi. In particolare, anche di recente, il legislatore specifica che l'interoperabilità avviene tra i sistemi di conservazione mediante la produzione di pacchetti di distribuzione coincidenti con i pacchetti di archiviazione o comunque contenenti pacchetti di archiviazione generati

struttura o al responsabile del servizio di conservazione in caso di affidamento eccetto la lettera m) del par. 4.5 delle stesse Linee Guida

[125] I metadati del documento informatico, del documento amministrativo informatico e della aggregazione documentale informatica sono indicati nell'Allegato 5 "Metadati" delle Linee guida

[126] Il glossario dei termini e degli acronimi contenuto nell'Allegato 1 "Glossario dei termini e degli acronimi" delle Linee guida definisce il rapporto di versamento il "*documento informatico che attesta l'avvenuta presa in carico da parte del sistema di conservazione dei pacchetti di versamento inviati dal produttore"*

[127] UNI, "Supporto all'Interoperabilità nella Conservazione e nel Recupero degli Oggetti digitali. SInCRO", N. 11386:2020, 2020

sulla base delle specifiche della struttura dati indicate dallo standard UNI 11386:2020. In tal senso, il nuovo Conservatore svolge la funzione di utente abilitato a richiedere l'accesso al sistema di conservazione del Conservatore che genera il PdD contenete il PdA e il relativo indice di conservazione. Secondo la logica del modello OAIS, ciò costringe il nuovo Conservatore a versare nuovamente il PdD all'interno del proprio sistema di conservazione dando vita a un processo oneroso che svaluta le potenzialità dell'interoperabilità tra i sistemi informativi. In tal senso, sarebbe auspicabile che si l'interoperabilità tra i sistemi di conservazione si incentrasse sullo scambio dei PdA per garantire una maggiore coerenza con il modello OIAS e valorizzare le potenzialità della interoperabilità. La recente pubblicazione del documento (dicembre 2022) elaborato dal secondo sottogruppo del più generale gruppo di lavoro sui Poli della conservazione relativo a "*Modelli di interoperabilità tra sistemi di conservazione*" mira a «individuare alcuni modelli di interoperabilità tra sistemi di conservazione partendo dalla definizione di un PdA interoperabile, in modo da garantire ai poli di conservazione lo scambio reciproco degli oggetti conservati, riducendo al minimo la perdita di informazioni»[128]. Il paragrafo "*Gli elementi per la definizione del modello di interoperabilità*" stabilisce che «i due principali strumenti individuati per definire il modello di interoperabilità sono:

- lo standard UNI 11386:2020 - Supporto all'Interoperabilità' nella Conservazione e nel Recupero degli Oggetti digitali (SInCRO)2 per la costruzione dell'Indice del PdA;
- l'allegato 5 alle LLGG, che contiene il set di metadati descrittivi».

In particolare, il par. 4.3. del citato documento chiarisce che la funzione del SiNCRO è la definizione degli aspetti strutturali e sintattici mentre quella dell'Allegato 5 è di natura semantica dell'unità documentaria. La soluzione proposta dal legislatore consiste nell'integrazione dello schema di metadati all'interno dell'elemento *MoreInfo* espresso nei vari livelli accettabili (*PVolume, FileGroup e File*) all'interno dello schema del SiNCRO. Il documento rappresenta indubbiamente un supporto essenziale per i soggetti coinvolti dal processo di conservazione sebbene lasci spazio di confronto. Per esempio, la mancata imposizione della tecnologia per la rappresentazione dei metadati potrebbe causare l'assenza degli appositi software da parte del Conservatore ricevente il PdD che rischierebbe di complicare la leggibilità e quindi la valorizzazione semantica del contenuto informativo[129]. Un'ulteriore attività del processo di conservazione consiste nella preparazione e sottoscrizione del PdD, secondo le modalità definite dalla let. g) del par. 4.7. delle Linee Guida. Tale azione si rende necessaria in caso di richiesta di esibizione da parte di un utente abilitato dal Titolare dell'oggetto di conservazione.

[128] Agenzia per l'Italia Digitale, "Modelli di interoperabilità tra sistemi di conservazione", Roma, 2021

[129] Agenzia per l'Italia Digitale, "FAQ", link: https://www.agid.gov.it/it/domande-frequenti/documento-informatico. Consultato il 27/02/2023

In particolare, il par. 4.9. delle Linee guida descrive le modalità di esibizione degli oggetto digitali e sottolinea che «il sistema di conservazione permette ai soggetti autorizzati l'accesso diretto, anche da remoto, agli oggetti digitali conservati, attraverso la produzione di pacchetti di distribuzione secondo le modalità descritte nel manuale di conservazione, prevedendo opportune misure tecniche e organizzative per garantire un livello di sicurezza adeguato al rischio e modalità di accesso diverse, in funzione delle tipologie di dati personali trattati, nonché delle operazioni di trattamento consentite».[130] Tra le misure di sicurezza un aspetto centrale è certamente occupato dal controllo degli accessi e del relativo rilascio delle credenziali agli utenti abilitati. Tali accessi devono essere concordati tra Titolare dell'oggetto di conservazione e Conservatore e definiti in rapporto alla normativa di riferimento, tra cui, per esempio, la L. del 7 agosto 1990 n. 241 in materia di procedimento amministrativo e di diritto di accesso ai documenti amministrativi e agli art. 33 e 34 del Regolamento europeo n. 679/2016 GDPR in materia di violazione dei dati personali. Inoltre, tra le misure di sicurezza in capo al responsabile della conservazione figura l'eventuale attività di riversamento dei formati al fine di garantire la leggibilità nel tempo in base alle indicazioni previste dall'allegato 2 "Formati di file e riversamento" e la consequenziale produzione di duplicati informatici o di copie informatiche effettuata anche su richiesta degli utenti. Infine, il processo di conservazione si conclude con la procedura di selezione per lo scarto o l'eventuale conservazione permanente degli oggetti digitali ritenuti di notevole valore storico descritta nel par. 4.11. delle Linee Guida. Tale procedura di scarto deve avvenire nel rispetto dei tempi stabiliti dalla norma di riferimento costituita dal Codice dei Beni culturali e del Paesaggio. Nel caso in cui il servizio di conservazione venga affidato ad un conservatore, parte o tutte le attività appena descritte, ad esclusione della predisposizione del manuale di conservazione potranno essere affidate al responsabile del servizio di conservazione. Tale affidamento dovrà essere formalizzato mediante accordo di affidamento del servizio di conservazione nel quale saranno indicati anche i nominativi ed i riferimento del responsabile della conservazione e del responsabile del servizio di conservazione e le relative attività delegate. La figura 4 illustra il modello di conservazione nel caso di affidamento all'esterno.

[130] • Agenzia per l'Italia Digitale, "Linee Guida sulla formazione, gestione e conservazione dei documenti informatici", link: https://www.agid.gov.it/sites/default/files/repository_files/linee_guida_sul_documento_informatico.pdf. Consultato il 27/02/2023

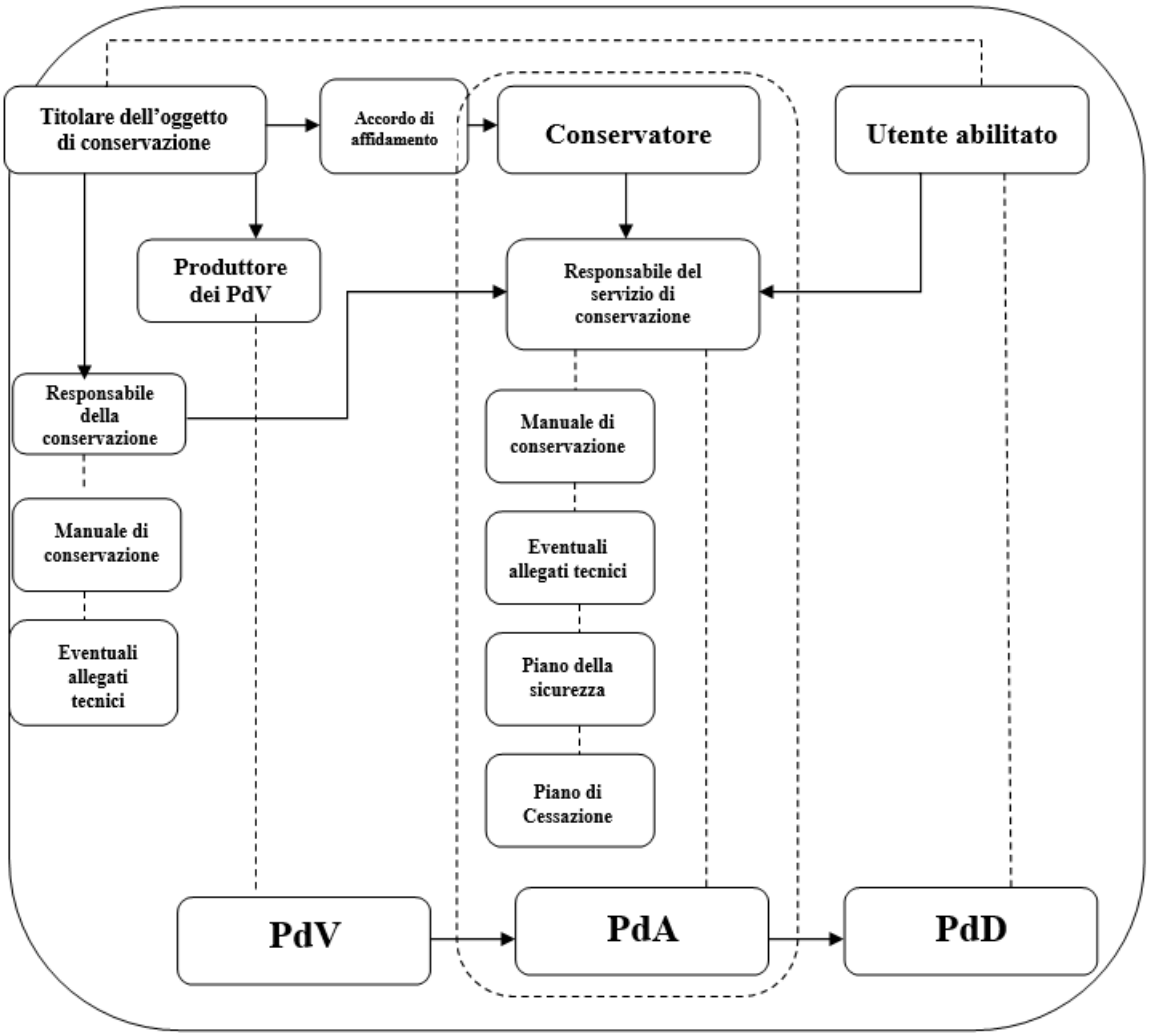

Figura 17. Il modello di conservazione in caso di affidamento all'esterno del servizio di conservazione

Il modello italiano della conservazione digitale costituisce certamente un aspetto essenziale per definire le modalità di conservazione degli oggetti digitali. Tuttavia, l'adozione di tale modello richiede una traduzione che sia funzionale rispetto al contesto di produzione delle risorse. Per esempio, nel caso dei prodotti della ricerca sarebbe opportuno definire cosa si intenda per prodotto della ricerca sottoposto a conservazione. Infatti, i prodotti della ricerca rappresentano una categoria estremamente omogenea di oggetti digitali che potrebbero riguardare i documenti nativi informatici o i documenti scannerizzati, i software o le basi di dati, i codici o i messaggi di posta elettronica. Un punto di partenza cruciale sarebbe pertanto individuare innanzitutto una precisa tipologia di oggetto digitale.

In rapporto all'oggetto digitale si rende necessario identificare il formato idoneo e il relativo schema di metadati da associare. In tal senso, l'oggetto digitale e i relativi metadati potrebbero comporre il pacchetto di versamento trasferito al sistema di conservazione, gestito al suo interno secondo le specifiche dei PdA e restituito sottoforma di PdD. Tuttavia, l'estrema semplificazione del processo appena descritto

nasconde diverse zone d'ombra che sarebbe necessario definire. Attualmente i prodotti della ricerca seguono un iter procedurali di conservazione differenti perché differenti sono le norme di riferimento[131]. Per esempio, i documenti amministrativi che danno avvio al procedimento e quindi all'attività di ricerca sono preservati nel tempo all'interno dei sistemi di conservazione, mentre i report, gli articoli, le basi di dati e i software sono archiviati in repository anche differenti. Ciò da un lato spezza il legame logico tra gli oggetti che testimoniano le attività di ricerca, limitando la possibilità di recupero e complicando la ricostruzione sequenziale dei documenti, e dall'altro affida la custodia della memoria scientifica a repository che non assicurano la leggibilità di tale patrimonio. Inoltre, sarebbe opportuno definire le responsabilità ricoperte dalle figure coinvolte nella produzione e conservazione di tale documentazione anche in rapporto ai ruoli previsti dalla normativa vigente anche in rapporto ai diritti di accesso da parte degli utenti abilitati. Infine, diviene indispensabile individuare le procedure di selezione e scarto in considerazione dei tempi stabiliti dagli organismi di controllo e formalizzati nei massimari di selezione e scarto di ciascun soggetto produttore in rapporto alla natura giuridica dello stesso e alla normativa di riferimento. In conclusione, la conservazione digitale dei prodotti della ricerca richiede una necessaria integrazione per la definizione di un modello che permetta di garantire il mantenimento delle caratteristiche di autenticità, affidabilità, integrità e della possibilità di reperire tali risorse nel tempo.

[131] Si veda il par. 1.4. del presente lavoro di tesi

2.1.2. Le metodologie e le tecniche di estrazione automatica di metadati

L'incremento esponenziale degli oggetti digitali e dei relativi metadati ad essi associati ha richiesto lo sviluppo di soluzioni idonee a migliorare l'efficace ed efficiente recupero di tali risorse informative. La compilazione manuale dei metadati è infatti ritenuta un'operazione eccessivamente onerosa che determina il maggiore rischio di incorrere in una erronea valorizzazione, compromettendo negativamente anche il recupero. Una delle soluzioni per mitigare tali rischi consiste nello sviluppo di sistemi di estrazione automatica di metadati. L'implementazione di tali sistemi può essere basata su diverse metodologie e tecnologie in rapporto all'oggetto digitale. Tra gli oggetti digitali ampiamente analizzati figurano i documenti testuali, sebbene nel tempo sono stati implementati sistemi in grado di estrarre metadati anche da oggetti multimediali (per esempio immagini o video). Per esempio, l'analisi della forma strutturale del documento testuale costituisce una delle metodologie ampiamente utilizzate anche grazie al supporto degli algoritmi di intelligenza artificiale. La scelta delle metodologie e tecnologie può dipendere dalle informazioni contenute, dai metadati obiettivo dell'estrazione e dal formato di input e output. Uno dei formati di input accettato dai sistemi di estrazione automatica di metadati da documenti testuali è il Portable Document Format (PDF), mentre generalmente l'output è fornito in XML (eXtensible Mark-up Language). In tal senso, la scelta delle metodologie e tecnologie è un passaggio fondamentale per rispondere adeguatamente al fabbisogno informativo necessario per garantire la conservazione e il recupero delle risorse. Pertanto, obiettivo del presente paragrafo è analizzare le principali strategie metodologiche e tecnologiche impiegate dai sistemi di estrazione automatica di metadati e infine illustrare le metriche di valutazione dei risultati ottenuti.

La principale metodologia impiegata dai sistemi di estrazione automatica di metadati da documenti testuali è incentrata sull'analisi del layout documentale (Document Layout Analysis). La *Document Layout Analysis* sfrutta l'osservazione e la collocazione delle informazioni all'interno del testo per individuare, elaborare ed infine estrarre la conoscenza strutturata. Tale processo prima comporta l'analisi della struttura fisica del layout (Physical Layout Analysis) e successivamente l'analisi della struttura logica (Logical Layout Analysis). Come sottolineato anche da Namboodiri et al. "the physical layout of a document refers to the physycal location and boundaries of various regions in the document image"[132] mentre la logical layout si definisce come "the set of logical or functional entities in a document, along with their inter-relationships is referred to as the Logical Structure of the Document. The analysis of logical structure of a document is usually performed on the results of the layout analysis stage"[133].

[132] NAMBOODIRI, Anoop M., JAIN, Anil K., "Document Structure and Layout Analysis", in *ResearchGate*, 2007. DOI: 10.1007/978-J-84628-726-8_2

[133] Ibidem

In generale, è possibile distinguere almeno tre approcci di Document Layout Analysis: (a) bottom-up, (b) top-down e (c) hybrid. L'approccio bottom-up (a) prima elabora gli elementi granulari del testo (es. i caratteri, le parole, le frasi e i paragrafi), poi calcola la distanza che intercorre tra essi fino a identificare zone di testo omogenee ed estrarne il contenuto. Al contrario, l'approccio top-down (b) prima riconosce le zone di testo, poi le divide in parti fino a giungere agli elementi del livello base. Infine, l'approccio ibrido (c) include sia l'approccio bottom-up sia l'approccio top-down al fine di migliorare la qualità in termini di precisione (qui intesa come corretta corrispondenza tra i metadati e i relativi valori) delle informazioni estratte. La scelta dell'approccio dipende principalmente dai metadati obiettivo dell'estrazione e dalle tecniche adoperate.

Le soluzioni per l'elaborazione delle informazioni sulla struttura fisica e logica sono incentrate sull'uso di algoritmi di intelligenza artificiale sviluppati su tecniche differenti in rapporto alla metodologia adottata. In generale, è possibile suddividere tali algoritmi in due macrocategorie: supervisionati e non-supervisionati. Per esempio, tra gli algoritmi supervisionati rientrano quelli sviluppati su tecniche di *Machine Learning* - che includono per esempio tecniche *Rule-based* o *Support Vector Machine*[134] - *Deep Learning* (per esempio Convolutional Neural Network – CNN) e *Natural Language Processing* mentre tra le tecniche adoperate per lo sviluppo di algoritmi non supervisionati rientra per esempio il *Clustering* (tra cui il K-Means). Eskenazi et al. ha proposta una tassonomia dei principali algoritmi utilizzati per la segmentazione dei documenti testuali.[135] Tale tassonomia suddivide gli algoritmi di segmentazione delle pagine dei documenti in tre gruppi secondo l'approccio top-down, bottom-up e ibrido e ciascun gruppo si compone di sottocategorie in rapporto alle tecniche adottate. Gli algoritmi del primo gruppo sono settati per segmentare una specifica tipologia di layout da cui estrarre le informazioni. Per esempio, in tale gruppo potrebbero rientrare gli algoritmi Rule-based. In tal senso, risulta utile in prima istanza identificare le tipologie di layout che potrebbe essere utilizzate per la formazione dei documenti testuali. Per esempio, Binmakhashen et al.[136] ha proposto una tassonomia non esaustiva dei principali stili utilizzati.

[134] HAN, Han, GILES, C.Lee, MANAVOGLU, Eren, ZHA, Hongyuan, ZHANG, Zhenyue, FOX, Edward A., "Automatic document metadata extraction using support vector machines", in *Joint Conference on Digital Libraries*, organized by ACM/IEEE, 2033. pp. 37–48

[135] ESKENAZI, Sébastien, KRÄMER, Petra Gomez, OGIER, Jean-Marc, "A comprehensive survey of mostly textual document segmentation algorithms since 2008", in *Pattern Recognition*, Elsevier, 2016. DOI: http://dx.doi.org/10.1016/j.patcog.2016.10.023

[136] BINMAKHASHEN, Galal M., MAHMOUD, Sabri A., "Document Layout Analysis: A comprehensive Survey", in *ACM Computing Surveys*, Vol. 52, No. 6, 2019;

Tale tassonomia prevede sei stili:

Figura 18. A) *Regular, b*) *Manhattan-based, c) non-Manhattan, d) Multi-column Manhattan, e) Arbitrary Complex, f) Overlapping horizontally and diagonally. (BINMAKHASHEN, Galal M., MAHMOUD, Sabri A., "Document Layout Analysis: A comprehensive Survey", in ACM Computing Surveys, Vol. 52, No. 6, 2019)*

Per esempio, Manhattan (b) e Multi-Column Manhattan (d) potrebbero rappresentare gli stili convenzionalmente utilizzati per i documenti testuali prodotti in ambito scientifico e analizzati dagli algoritmi del primo gruppo. L'adeguata individuazione del layout costituisce un passaggio essenziale che permette di migliorare l'identificazione delle informazioni contenute nel documento arricchendo i metadati associati alla risorsa con ricadute positive anche in termini di qualità nel recupero delle informazioni.

L'adozione arbitraria del layout, anche per la medesima tipologia documentale prodotto nello stesso contesto applicativo, e l'assenza di standard che indichino le specifiche tecniche per la creazione di tale layout genera alcuni limiti nei sistemi di elaborazione e segmentazione delle pagine. Inoltre, i documenti potrebbero essere composti da altri oggetti informativi oltre al testo (per esempio tabelle, grafici e immagini). Ciò richiederebbe una maggiore accuratezza nella corretta individuazione anche di tali oggetti. Per fornire una risposta a tali riflessioni sono stati sviluppati algoritmi in grado di adattarsi alle variazioni della struttura del documento per poter segmentare una gamma più ampia di layout con lo stesso algoritmo. Tali algoritmi sono quelli che compongo il secondo gruppo, tra cui gli algoritmi di clustering, quelli basati sulle analisi funzionali e quelli di classificazione addestrati a riconoscere il tipo di elemento. Infine, gli algoritmi del terzo gruppo combinano i sistemi di analisi bottom-up e top-down. Tra questi algoritmi rientrano per esempio quelli incentrati sulle tecniche di deep learning come gli algoritmi neural network. La capacità di combinare gli algoritmi del primo e del secondo gruppo permette di ottenere risultati più complessi rispetto al solo utilizzo dell'uno o dell'altro gruppo. Tuttavia, ciò richiede un

maggior dispendio di tempo per l'elaborazione e non sempre i risultati attesi sono assicurati.

In considerazione di ciò, secondo Golub et al. ha tentato di atomizzare gli step di analisi individuando almeno cinque modalità per l'elaborazione delle informazioni anche in rapporto al formato del documento :1) l'analisi della formattazione documentale (Formatting Structure), 2) l'analisi della strutturale visuale (Visual Structure), 3) l'analisi del layout documentale (Document Layout), 4) l'analisi delle citazioni bibliografiche (Bibliographic citation analysis) e infine 5) l'analisi della struttura linguistica (Linguistic structure)[137].

In particolare, la 1) permette di estrarre la struttura da documenti formati attraverso i linguaggi di mark-up come HTML (HyperText Markup Language)[138]. Il vantaggio dell'estrazione di informazioni strutturali da documenti HTML consiste nell'ottenere con alta precisione la corretta posizione delle informazioni definita dai tag (per esempio H1, H2 o H3 sono i tag che indicano l'inizio dell'intestazione). Una ricaduta positiva in merito all'applicazione di tale procedura di estrazione potrebbe consistere nella conservazione delle pagine web che costituiscono un importante strumento di trasmissione conoscitiva globale e che rischiano quotidianamente di scomparire (si pensi alla chiusura di un blog di divulgazione scientifica). Tuttavia, raramente i documenti sono redatti in HTML e ciò ne limita l'utilizzo.

La 2) e la 3) sono entrambe modalità utilizzate per l'estrazione della struttura da documenti in formato PDF. La prima sfrutta i metadati strutturali che indicano la disposizione delle informazioni in una pagina mentre la seconda si basa sull'individuazione delle informazioni nelle zone del testo in rapporto alla convenzione grafica adoperata per la tipologia documentale (per esempio gli articoli scientifici o i report della ricerca). La propedeuticità essenziale di tali modalità ha permesso di raggiungere risultati significativi nel campo dell'estrazione automatica di metadati da documenti nativi digitali anche grazie alla capacità di elaborazione di formati di input e output, ampiamente diffusi, come il PDF e l'XML.

La modalità 4) è finalizzata all'estrazione delle informazioni citazionali. Tali informazioni sono particolarmente utili ai ricercatori per acquisire contributi scientifici al fine di definire lo stato dell'arte del proprio campo di ricerca. Inoltre, estrarre le informazioni citazionali permette di legare gli argomenti trattati ampliando i confini della conoscenza della ricerca e ipotizzare anche eventuali nuovi scenari. L'individuazione della sezione citazionale potrebbe essere affrontata anche attraverso la metodologia DLA grazie alla convenzionale collocazione al termine di ciascuna pagina o al termine dell'intero testo. Successivamente, l'elaborazione del contenuto e l'estrazione (attraverso le tecniche indicate nel paragrafo di riferimento) strutturano

[137] GOLUB, Koraljka, MULLER, Henk, TONKIN, Emma, "Technologies for Metadata Extraction", in *Semantics and Ontologies*, World Scientific, 2014. ISBN: 9789812836298

[138] W3Schools, "HTML", link: https://www.w3schools.com/html/. Consultato il 27/02/2023

finalmente le informazioni. La maggior parte dei risultati sono attualmente formalizzati in XML, tuttavia, l'uso di formati come RDF contribuirebbe a collegare ulteriormente il patrimonio conoscitivo attraverso la definizione di *knowledge graph*.

Infine, la modalità 5) permette l'analisi linguistica (Linguistic structure)[139] del contenuto testuale. Tale analisi è spesso incentrata sull'utilizzo di tecniche in grado di elaborare la struttura sintattica dei periodi la cui formulazione dipende dalla lingua adoperata e dalle relative regole grammaticali.

In tal senso, il primo passo per ottenere risultati significativi dall'analisi linguistica è identificare la tipologia di lingua adoperata per la produzione documentale. Le informazioni estratte possono variare in rapporto all'obiettivo dell'estrazione e tipicamente includono concetti-chiave, agenti, eventi o luoghi descritti nel testo. La funzione dell'analisi linguistica permette quindi la descrizione sintetica del contenuto del documento, utile ai fini del recupero, ma anche di valorizzare semanticamente il contenuto informativo contestualizzando e relazionando i termini in rapporto al dominio applicativo.

Obiettivo finale dell'applicazione di tali algoritmi resta l'analisi della struttura fisica del layout e quella della struttura logica del documento.

Le immagini seguenti mostrano a titolo esemplificativo prima il risultato ottenuto dalla prima analisi e dopo quello ottenuto dalla seconda.

[139] GOLUB, Koraljka, MULLER, Henk. TONKIN, Emma, "Technologies for Metadata Extraction", in *Semantics and Ontologies*, World Scientific, 2014. ISBN: 9789812836298

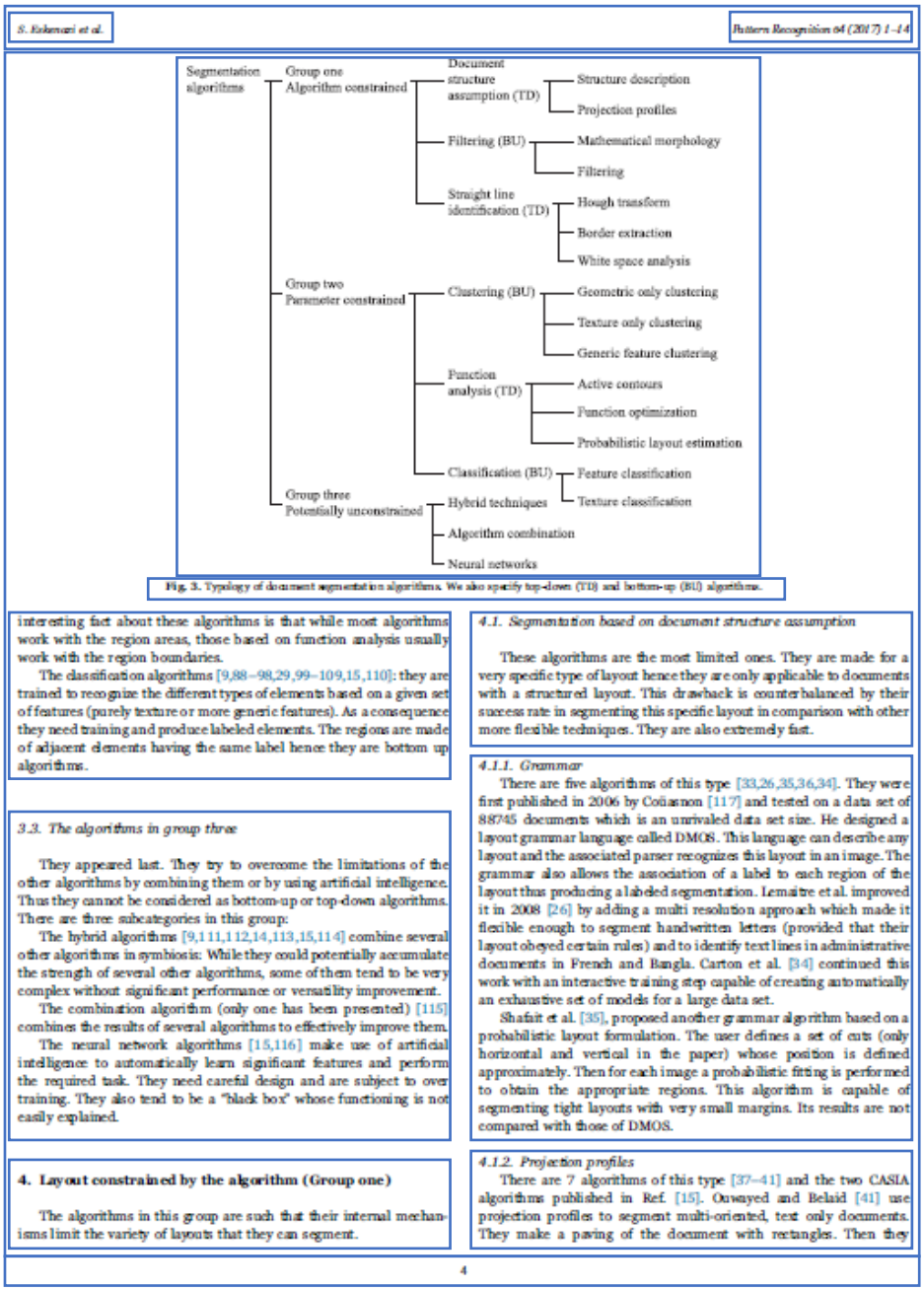

Figura 19. Esempio di risultato ottenuto al termine dell'analisi della struttura del layout di un documento

L'analisi della struttura del layout documentale permette di identificare la posizione delle informazioni nelle aree del testo. Generalmente, il processo di identificazione è basato sul calcolo della distanza tra i caratteri attraverso, per esempio, l'uso delle tecniche di clustering o quelle rule-based. Tuttavia, tale risultato potrebbe non essere sempre garantito. Per esempio, l'estrema vicinanza tra i caratteri potrebbe impedire la distinzione tra paragrafi oppure la distanza ridotta tra i caratteri potrebbe comportare l'inclusione al medesimo paragrafo composto in realtà da paragrafi differenti.

L'errato disconoscimento della appartenenza degli oggetti informativi del documento a una zona determina certamente ricadute negative nella seconda fase della logical layout analysis compromettendo l'intero processo di estrazione automatica di metadati.

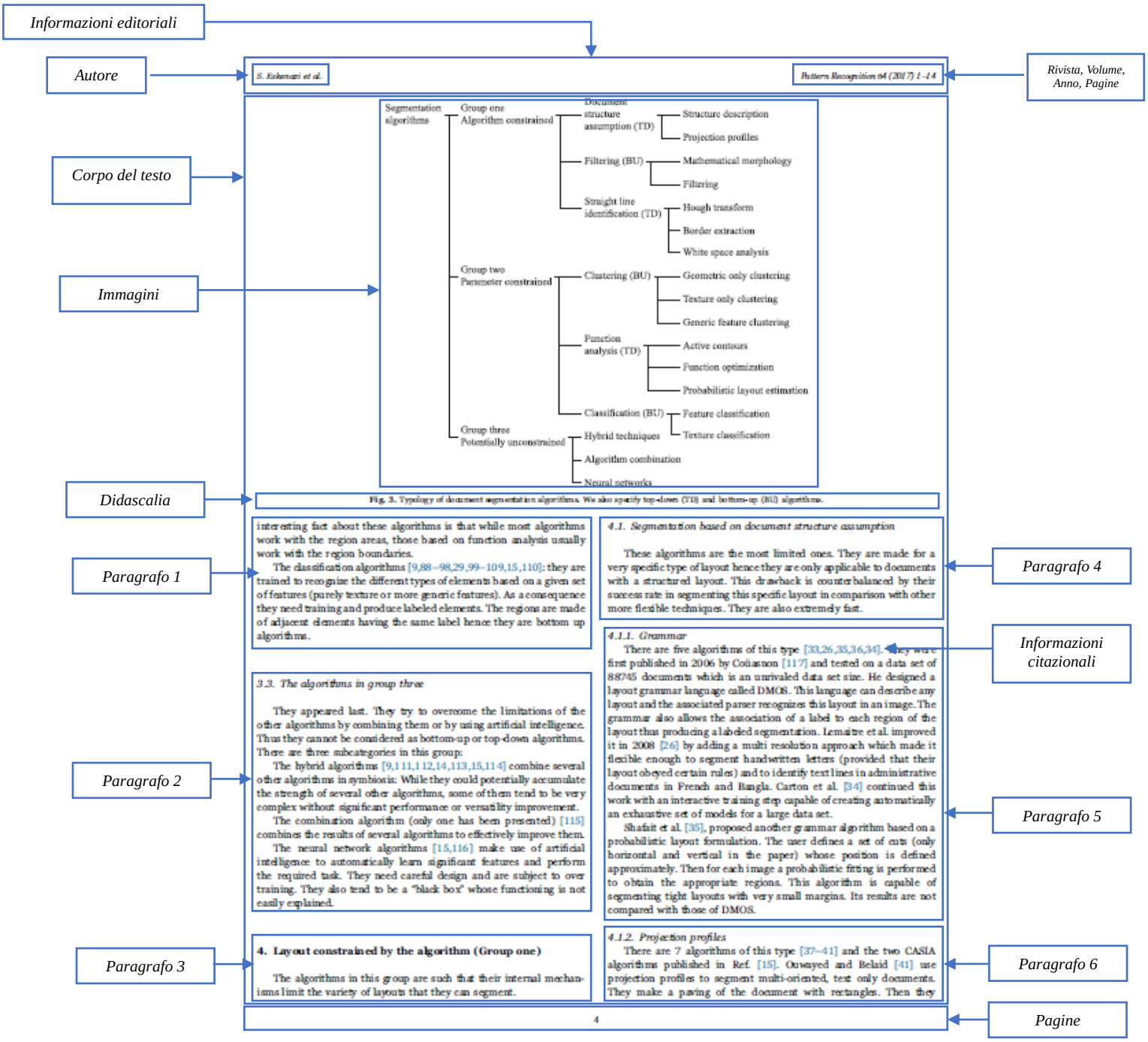

Figura 20. Esempio di associazione dei metadati in rapporto alla zona di testo

Le zone precedentemente identificate grazie alla posizione degli elementi sono ora distinte per funzioni o per entità durante l'analisi della struttura logica. L'analisi della struttura logica del layout è tendenzialmente basata su algoritmi di machine learning, tra cui quelli di classificazione, come le SVM (Support Vector Machine). In particolare,

la classificazione delle zone di testo permette l'identificazione delle funzioni dal punto di vista informativo rivestite da tali zone. Ciò costituisce un passaggio essenziale per individuare i possibili metadati estraibili. Inoltre, l'applicazione di tali algoritmi richiede una necessaria fase di training per la definizione del modello da analizzare ed è pertanto essenziale durante tale fase individuare la tipologia di layout e di informazioni da analizzare. In tal senso, è cruciale che la fase di training sia supervisionata da un esperto di dominio al fine di identificare i potenziali metadati appartenenti al contesto applicativo e che potrebbero essere estratti dal documento.

Le informazioni identificate al termine dell'analisi sono estratte e strutturate secondo le specifiche dei metadati adottati e restituiti in formati idonei a garantire la leggibilità e l'interoperabilità tra i sistemi. Dal punto di vista progettuale, la maggior parte delle soluzioni prevedono l'uso combinato di tali modalità. Tale combinazione dovrebbe migliorare la ricchezza e la qualità delle informazioni estratte.

Criteri di valutazione dei risultati

Nonostante i significativi risultati raggiunti nel campo dell'estrazione automatica di metadati vi sono ancora ampi spazi di confronto e miglioramento. Infatti, non è sufficiente estrarre metadati ma è necessario verificare la corretta o incorretta corrispondenza tra il metadato e il valore informativo. In particolare, l'errata corrispondenza può riguardare o la mancata estrazione del valore oppure la incorretta corrispondenza del valore con il metadato. Il primo caso potrebbe dipendere dal mancato riconoscimento dell'informazione oppure dall'associazione di quel valore ad un metadato differente. Il secondo caso potrebbe dipendere dall'estrema vicinanza dei caratteri che determina l'incorretto riconoscimento delle zone di testo e quindi delle relative informazioni. Entrambi gli errori sarebbero generati a monte del processo di estrazione fin dalla fase di structural analysis. In tal senso, per valutare la completezza e corretta corrispondenza semantica tra i metadati e i relativi valori si rende necessaria una fase di valutazione dei risultati incentrata sulla misurazione e successiva analisi. Tra le metriche di misurazione attualmente impiegate figurano per esempio la Precision e la Recall. La precision può essere definita come il rapporto tra numero e numero totale di metadati, mentre la recall consiste nel rapporto tra il numero di metadati estratti e il numero di metadati correttamente valorizzati. Infine, un'ulteriore metrica di valutazione è l'F-Measure che combina entrambe le metriche di Precision e Recall all'interno di un singolo sistema di misurazione. In particolare, la F-measure assume valori compresi tra 0-e 1 in cui 0 corrisponde a un valore molto basso, 1 a un valore molto alto. Per esempio, l'F-measure assume valore 0 quando non sono stati estratti metadati rilevanti e 1 quando sono stati estratti tutti i metadati rilevanti. Le metriche di valutazione permettono di validare la correttezza e la completezza dei metadati estratti e delle relative informazioni, anche grazie alla partecipazione di esperti archivisti. Se l'utilizzo delle metriche costituisce una soluzione alla problematica gestione della valutazione dei risultati, attualmente ignorata dai sistemi di estrazione automatica di metadati, restano ancora aperti ulteriori campi di discussione, come quello legato alla

scelta del formato adottato. Per esempio, il PDF è uno dei formati più adottato per la produzione di documenti. Il principale vantaggio offerto dall'adozione del PDF è l'apertura delle specifiche tecniche che lo rende indipendente da hardware e software aumentando la probabilità di leggibilità nel tempo). Tuttavia, "most PDF documents are untagged and don't have the basic high level logical structure information such as words, text lines, paragraphs, logos and figure illustrations, which makes reusing, editing or modifying the layout or the content of the document difficult".[140] Pertanto, è necessario identificare ed estrarre le informazioni sulla struttura fisica e logica da un documento in formato PDF ma ciò non è immediato. Inoltre, la maggior parte delle soluzioni tendono a fornire uno schema minimo di metadati che non soddisfa i requisiti necessari per garantire la conservazione e il recupero dei documenti nel tempo. Inoltre, gli schemi di metadati sono spesso definiti arbitrariamente dagli stakeholders. La definizione arbitraria di schemi di metadati non basati su standard internazionali o sulla base delle indicazioni fornite dalla normativa di riferimento determina una scarsa interoperabilità tra i sistemi, un maggiore rischio di irreperibilità dei documenti con ricadute negative anche in termini economici. In tale ottica, si rende necessario adattare gli schemi di metadati a quelli previsti dal quadro di riferimento nazionale in merito alla conservazione dei documenti informatici. Pertanto, il prossimo paragrafo intende analizzare il modello della conservazione digitale adottato nel contesto nazionale anche in considerazione delle politiche di metadatazione.

[140] ALTO, Palo, "Layout and Content Extraction for PDF Documents". In *Springer*, *Verlag Berling Heidelberg*, 2004. pp. 213-224

2.2. Dalla base di conoscenza al framework

L'analisi della base di conoscenza rappresenta un passaggio preliminare per valutare le informazioni minime essenziali che il framework di estrazione automatica di metadati per la conservazione dei prodotti della ricerca deve essere in grado di elaborare. La valutazione delle informazioni influenza, infatti, le fasi successive incentrate sulla modellazione concettuale e lo sviluppo del sistema determinando anche la qualità dei risultati finali. Prima di approfondire tale tematica è necessario, tuttavia, fornire una definizione relativa al significato del termine "base di conoscenza", variabile in rapporto al contesto applicativo. Per esempio, nell'ingegneria informatica una "base di conoscenza" indica una base di dati prodotti da una specifica organizzazione il cui accesso è limitato a un ristretto numero di individui in rapporto ai diritti posseduti. Secondo Hess e Ostrom "tutte le forme di sapere conseguito attraverso l'esperienza o lo studio, sia esso espresso in forma di cultura locale, scientifica, erudita o in qualsiasi altra" [141] sono entità egualmente accessibili che trasmettono conoscenza alla comunità di riferimento. In questo contesto impieghiamo la nozione di "base di conoscenza" per riferirci al complesso di oggetti digitali che compongono la categoria dei prodotti della ricerca. Pertanto, l'obiettivo finale del presente paragrafo consiste nella descrizione della base di conoscenza del framework mediante un duplice approccio:

- ✓ identificare il complesso di oggetti digitali afferenti alla categoria dei prodotti della ricerca;
- ✓ definire le caratteristiche di tali oggetti digitali.

Sebbene i prodotti della ricerca sia un complesso di entità noto all'universo scientifico, e non solo, la loro identificazione non è affatto istantanea. Infatti, l'assenza di un quadro istituzionale che identifichi definitivamente le tipologie di oggetti che compongono i prodotti della ricerca ingenera spesso confusione anche nella stessa comunità di riferimento. Tale confusione è accentuata poi dalla eterogeneità dei repository implementati per ciascuna area di ricerca spesso contenenti le medesime tipologie documentali ma denominate e organizzate in modo differente[142].

Convenzionalmente, i prodotti della ricerca sono distinti in:

1. Prodotti della ricerca pubblicati attraverso gli ordinari canali commerciali;
2. Prodotti della ricerca non pubblicati attraverso gli ordinari canali commerciali.

In particolare, la classe dei prodotti della ricerca pubblicati attraverso gli ordinari canali commerciali può essere suddivisa in due ulteriori sottoclassi:

[141] a cura di HESS, Charlotte, OSTROM., Elinor, La conoscenza come bene comune. Dalla teoria alla pratica, Bruno Mondadori, Milano, 2009. ISBN: 978-88-6159-142-4

[142] Commissione Europea, "Project Databases", link: https://research-and-innovation.ec.europa.eu/projects/project-databases_en. Consultato il 27/02/2023

1.1. Letteratura bianca;
1.2. Letteratura nera.

La sottoclasse afferente alla letteratura bianca racchiude oggetti digitali pubblicati attraverso gli ordinari canali commerciali che, generalmente, include i prodotti divulgativi dell'attività di ricerca la cui pubblicazione può essere gratuita o a pagamento. La sottoclasse della letteratura nera include al contrario tipologie di prodotti inaccessibili poiché ritenuti riservati.

La seconda classe è relativa ai prodotti della ricerca non pubblicati attraverso gli ordinari canali commerciali nota come:

1.3. Letteratura Grigia.

Come evidenziato da *Metitieri e Ridi* «il riferimento al colore grigio è nato a metà degli anni Settanta e allude a qualcosa di intermedio fra la normale letteratura «bianca» dei circuiti commerciali e quella «nera», completamente inaccessibile. Questa grande categoria di documenti include fra l'altro tesi di laurea e di dottorato, rapporti tecnici aziendali, relazioni presentate a convegni, saggi in attesa di accettazione da parte di periodici accademici, cataloghi e manuali di prodotti hardware e software, dispense universitarie e relative a corsi di formazione. La letteratura grigia, in effetti, ha confini sfumati, tanto che qualcuno la estende fino a coprire tutti i documenti che non rientrano nelle tradizionali categorie delle monografie e dei periodici»[143].

Il modello convenzionalmente utilizzato per l'organizzazione dei prodotti della ricerca non fornisce, tuttavia, un quadro esaustivo degli oggetti interessati ma, per alcuni versi, contribuisce ad alimentare ulteriormente la già diffusa opacità.

Una soluzione utile per l'identificazione dei prodotti della ricerca può essere incentrata sull'analisi degli oggetti archiviati presso l'archivio ZENODO. Gestito da OpenAIRE e implementato grazie ai fondi dell'UE, ZENODO fornisce infatti indicazioni trasparenti sulla quantità e sulle tipologie di prodotti della ricerca depositati (che attualmente si attesta all'incirca su 2.845.859[144]).

In particolare, tali oggetti sono attualmente suddivisi secondo una organizzazione tassonomica di seguito illustrata:

Classe	**Tipologia**	**Numero**
Pubblicazioni	Articoli	1074151
	Trattamenti	451141

[143] METITIERI, Fabio, RIDI, Riccardo, Biblioteche in rete. Istruzioni per l'uso, Laterza, 2007. EAN: 9788842076636
[144] OpenAIRE, "ZENODO", link: https://zenodo.org/. Consultato il 27/02/2023

	tassonomici	
	Conference paper	75471
	Report	18283
	Altro	17746
	Deliverables	12933
	Libri	12747
	Sezioni	11684
	Tesi	8197
	Preprint	6888
	Working paper	5270
	Note tecniche	2480
	Proposte	692
	Documentazione Software	535
	Milestone	467
	Data Management Plan	365
	Peer review	322
	Brevetti	265
	Collezione di annotazioni	201
		Tot. **1699737+**
Immagini	Figure	501385
	Foto	280667
	Drawing	4518
	Altro	1114
	Plot	289
	Diagrammi	255
		Tot. **788228 +**
Dataset		Tot. **179830**
Software		Tot. **94946**
Presentazioni		Tot. **37084**
Altro[145]		Tot. **17735**
Poster		Tot. **15561**

[145] La categoria "Altro" può includere prodotti della ricerca non identificabile con le altre tipologie come per esempio basi di dati, tool o workflow

Video		Tot. **7402**
Lezioni		Tot. **4391**
Physicalobject		Tot. **945**
Tot.		**2.845.859**

Figura 21. Organizzazione delle tipologie di prodotti della ricerca in ZENODO

Il trasferimento in ZENODO può avvenire per almeno due ragioni:

1. L'obbligatoria disseminazione dei risultati ottenuti al termine di tutti i progetti finanziati da Horizon 2020 e Horizon Europe attraverso il deposito in accesso aperto in repository che rispettino i requisiti OpenAIRE (come ZENODO);
2. La condivisione aperta dei risultati scientifici da parte dei ricercatori il cui istituto di affiliazione è privo di un repository istituzionale[146].

Un ulteriore strumento per l'identificazione dei prodotti della ricerca è CORDIS «la principale fonte della Commissione europea in merito ai risultati dei progetti finanziati dai programmi quadro dell'UE per la ricerca e innovazione, dal 1º PQ a Horizon Europe»[147]. In particolare, CORDIS contiene la maggior parte dei progetti finanziati dai programmi realizzati dall'UE dal 1990:

- Progetti del 7° PQ;
- Horizon Europe;
- Horizon 2020.

Ciascun progetto si compone di:

1. Scheda informativa;
2. Risultati in breve;
3. Segnalazione;
4. Risultati;
5. Notizie e multimedia.

Per esempio, 1) la "Scheda informativa" contiene le informazioni descrittive relative al progetto come, per esempio:

- l'ID della convenzione;

[146] Non è possibile indicare in modo esaustivo lo spazio temporale dei prodotti condivisi su ZENODO in quanto non tutti dispongono di una data o, in alcuni casi, è stata associata una data inesatta (per esempio "1 Jen 9999").
[147] Commissione Europea, "CORDIS", link: https://cordis.europa.eu/about/it. Consultato il 27/02/2023

- la data di inizio;
- lo stato di avanzamento e l'eventuale data di fine;
- la denominazione del soggetto a cui è stato finanziato il progetto;
- il costo totale;
- i campi di ricerca;
- le keywords;
- l'obiettivo del progetto.

Tuttavia, la scheda può contenere al suo interno anche ulteriori tipologie documentali come:

- la richiesta di proposte;
- lo schema di finanziamento.

I 2) "Risultati finali" contengono informazioni sintetiche sui risultati ottenuti al termine dell'attività di ricerca mentre le 3) "Segnalazioni" svolgono la funzione di reporting in merito alle attività di ricerca e i relativi risultati prodotti in un arco temporale. I 4) "Risultati" contengono i prodotti finali della ricerca suddivisibili tra:

- "Documenti e report" (come, per esempio, documenti amministrativi, deliverables o reports tecnici);
- "Open Research Pilot" (per esempio Data management Plan);
- "Pubblicazioni" (tra cui articoli revisionati, capitoli, tesi di dottorato, monografie o atti di convegno);
- "Websites, brevetti e video".

Infine, le 5) "Notizie e multimedia" include indicazioni in merito ad eventi organizzati per la presentazione del progetto.

La tabella mostra l'organizzazione tassonomica dei prodotti della ricerca trasferiti CORDIS.

Classe	**Sottoclasse**		**Tipologia**
Scheda informativa			Richiesta di proposta
			Scheda di finanziamento
Risultati finali			Risultati finali
Segnalazioni			Report periodici
Risultati	Prodotti finali	Documenti e rapporti	Documenti amministrativi
			Deliverables
			Reports tecnici
			Relazioni
			Manuali

			Database
		Websites, brevetti e video	Websites
			Brevetti
			Video
		Open Research Data Pilot	Data management Plan
		Altro	Deliverabels di workshop
		Dimostrazioni, Prototipi	Deliverables di dimostrazioni
			Prototipi
	Set di Dati		Dataset
	Pubblicazioni		Articoli
			Capitoli
			Tesi di dottorato
			Monografie
			Atti di convegno
	Software		Software
Notizie e multimedia			Websites

Figura 22. Organizzazione delle tipologie di prodotti della ricerca in CORDIS

L'analisi degli oggetti depositati nei repository ZENODO e CORDIS permette quindi l'identificazione complessiva degli oggetti digitali che compongo la categoria dei prodotti della ricerca e quindi della base di conoscenza. Tale analisi mette in luce anche lo stretto rapporto che esiste, nel contesto della ricerca, tra i documenti e le risorse bibliografiche, afferenti a istituzioni e modelli differenti che, tuttavia, a volte rischiano di sovrapporsi. Offrire una definizione del termine documento potrebbe aiutare a chiarirne il significato e le ragioni della distinzione tra gli oggetti digitali. Tuttavia, come nel caso dei prodotti della ricerca, anche per i documenti, fornire una definizione univocamente condivisa non è immediato. In questo contesto ci si limita a sottolineare l'asserzione avanzata da Paola Carucci secondo cui il documento è il «prodotto nell'esercizio dell'attività pratica di una persona o di un ente, [e] costituisce la rappresentazione formale di un atto giuridico o di un fatto, caratterizzata dalla inscindibilità di forma e contenuto ovvero dalla staticità della sua natura»[148].

La definizione di documento della Carucci sottolinea il rapporto tra la funzione del soggetto produttore e il documento inteso come attestazione dell'avvenimento di un fatto giuridico realizzato nel rispetto di tale funzione. In virtù di tale rapporto il soggetto produttore forma quindi i documenti afferenti a un medesimo affare o pratica in rapporto alle funzioni svolte e legati da un vincolo necessario e spontaneo, di contenuto e di competenza. Tale vincolo e le relative caratteristiche sono ciò che

[148] CARUCCI, Paola, GUERCIO, Maria, *Manuale di Archivistica. Nuova Edizione*, Roma, Carocci Editore, 2021, p. 384

distinguono l'archivio dalla biblioteca in cui le risorse bibliografiche sono organizzate secondo la volontà dell'Istituto.

Tra gli esempi di prodotti della ricerca con valore documentale troviamo le proposte di ricerca, la pianificazione per la gestione dei finanziamenti, report, relazioni e basi di dati, mentre tra gli esempi di prodotti della ricerca intesi come risorse bibliografiche figurano la letteratura scientifica come gli articoli, gli atti di convengo, le monografie e le tesi di dottorato.

Valutare le caratteristiche di tali tipologie documentali è fondamentale affinché il framework rispetti i requisiti necessari per garantire il raggiungimento dell'obiettivo finale, cioè l'estrazione automatica di metadati e la conservazione dei documenti, poiché, alcune di queste, potrebbero influire sul processo di modellazione concettuale e infine di sviluppo dello stesso framework. La principale caratteristica comune alla maggior parte dei documenti afferenti ai prodotti della ricerca è la natura digitale. Ciò che varia è la modalità con cui tali documenti possono essere formati. Per esempio, tali documenti potrebbero essere formati mediante le seguenti modalità, in linea con le Linee Guida di AgID:

a) creazione tramite l'utilizzo di strumenti software;
b) acquisizione di copia per immagine su supporto informatico di un documento originale analogico;
c) memorizzazione su supporto informatico in formato digitale delle informazioni risultanti da transazioni o processi informatici o dalla presentazione telematica di dati attraverso moduli o formulari resi disponibili all'utente;
d) generazione o raggruppamento anche in via automatica di un insieme di dati o registrazioni, provenienti da una o più banche dati, anche appartenenti a più soggetti interoperanti, secondo una struttura logica predeterminata e memorizzata in forma statica.

I documenti possono quindi essere distinti in documenti di natura statica e documenti di natura dinamica[149]. La distinzione deriva dalla tassonomia fornita dagli autori del progetto InterPARES 2[150] secondo cui i documenti possono appunto essere distinti in documenti statici e interattivi, che a loro volta possono essere suddivisi in dinamici e non dinamici. Ciascuna delle classi documentali è incentrata su caratteristiche che richiedono l'adozione di misure adeguate per la conservazione a lungo termine che garantiscano il mantenimento dei requisiti di autenticità, di affidabilità, di accessibilità e la possibilità di recupero fin dalla fase di formazione. Per

[149] Un esempio di un documento di natura statica potrebbe essere un report elaborato tramite software con l'apposizione di una firma elettronica mentre un esempio di documento dinamico potrebbe essere una base di dati o un sito web.

[150] InterPARES 2, "The Nature of the Records and of the Processes that Create and Maintain Them" http://www.interpares.org/ip2/ip2_domain1.cfm . Consultato il 27/02/2023

esempio, tra i requisiti che è necessario definire vi sono l'associazione di uno schema di metadati e la scelta dei formati. Attualmente, lo schema associato ai prodotti della ricerca è incentrato su un insieme di metadati minimi, principalmente descrittivi, anche definiti sulla base di standard internazionali (come il Dublin Core) che non assicura la gestione dell'interno ciclo di vita dei documenti, mentre i principali formati adoperati sono l'**.HTML, *.PDF, *.PNG, *.JPG, *.ZIP, *.XLSX, *.XML, *.DOCX, *.TXT, *.MP4, *.CSV e *JSON* sebbene possano essere anche impiegati formati tecnologicamente obsoleti.

Infine, la conservazione costituisce l'esigenza cruciale di qualunque organizzazione che garantisce l'accessibilità ai documenti. Tuttavia, l'affidamento dei documenti a repository non è sinonimo di conservazione e non garantisce necessariamente la certezza della accessibilità nel tempo. Inoltre, se l'obbligo di disseminazione delle pubblicazioni ad accesso aperto prodotte al termine di progetti finanziati dai programmi quadro dell'Unione Europea Horizon 2020 e Horizon Europe pone un primo parziale argine alla complicata questione della gestione dei prodotti della ricerca, non è tuttavia sufficiente a garantire la certezza della conservazione e dell'accessibilità dei documenti nel tempo. Tale rischio deriva da almeno tre osservazioni:

1. Non è obbligatorio pubblicare tutti i documenti prodotti da attività di ricerca finanziata dai programmi europei;
2. Non tutti i documenti prodotti da attività di ricerca finanziata dai programmi europei sono pubblicati dai ricercatori;
3. i progetti di ricerca finanziati con fondi pubblici non derivanti dai programmi europei producono anch'essi documenti che tuttavia non è obbligatorio conservare.

Per esempio, sarebbe opportuno che il processo di conservazione si basasse su regole elaborate su standard internazionali e formalizzate all'interno di documenti tecnici appositamente definiti.

Tali criticità è accentuata poi nel caso di conservazione digitale di documenti dinamici con valore documentale. Per esempio, la conservazione di una base di dati afferente ai progetti di ricerca potrebbe essere incentrata su almeno tre strategie, tra cui l'*emulazione*, la *migrazione* e l'*hardware museum.* La scelta dipende da diversi fattori, tra cui la complessità della base di dati e l'obiettivo della conservazione. Sebbene alcuni progetti abbiano sperimentato la conservazione delle basi di dati[151], la scarsa attenzione prestata a tale tema, testimoniata anche dalla esigua letteratura, e i risultati ottenuti lasciano ampio margine di confronto.

In conclusione, la modellazione concettuale del framework di estrazione automatica di metadati per la conservazione dei prodotti della ricerca deve tener presente le

[151] Si pensi, per esempio al progetto RODA per la conservazione delle basi di dati delle pubbliche amministrazioni portoghesi incentrate sul formato SIARD e gestite secondo le logiche dell'ISO 14721:2012 OAIS

tipologie di oggetti identificati le relative caratteristiche che insieme definiscono la base di conoscenza qui analizzata che include diversi aspetti, tra cui le politiche, le metodologie, le tecnologie e gli output forniti. Tale modellazione deve quindi prevedere soluzioni migliorative che assicurino un'adeguata estrazione automatica di uno ricco set di metadati direttamente dal contenuto dei documenti e che garantisca la gestione del ciclo di vita dei documenti fino alla conservazione in sistemi appositamente implementanti per garantire l'autenticità, l'affidabilità, l'integrità e il recupero

2.2.1. La modellazione concettuale

Lo sviluppo di qualunque sistema richiede la realizzazione preliminare del relativo modello concettuale. Il modello concettuale può essere considerato un complesso di conoscenza esplicita che mira a descrivere gli elementi, le caratteristiche e le varie relazioni che ne intercorrono al fine di supportare la fase finale di sviluppo.

Esiste quindi una stretta dipendenza tra la conoscenza prodotta all'interno di un contesto applicativo e il relativo modello concettuale. Il presente paragrafo intende illustrare il modello concettuale del framework di estrazione automatica di metadati per la conservazione digitale dei prodotti della ricerca. L'intenzione finale è quella di proporre un modello che raggiunga l'obiettivo prefissato dal lavoro di ricerca, attraverso l'adozione di maggiori funzionalità rispetto alle attuali soluzioni, garantendo costi e tempi di realizzazione ridotti.

Prima di illustrare il modello concettuale è necessario analizzare la metodologia utilizzata per la definizione del modello. In particolare, esistono differenti metodologie in rapporto al contesto e all'obiettivo definito. La metodologia di lavoro qui impiegata per la modellazione concettuale del framework è basata su un approccio bottom-up che prevede quattro fasi sequenziali in cui l'output prodotto da una fase costituisce l'input della fase successiva. La figura 1 illustra le quattro fasi per la realizzazione del modello concettuale:

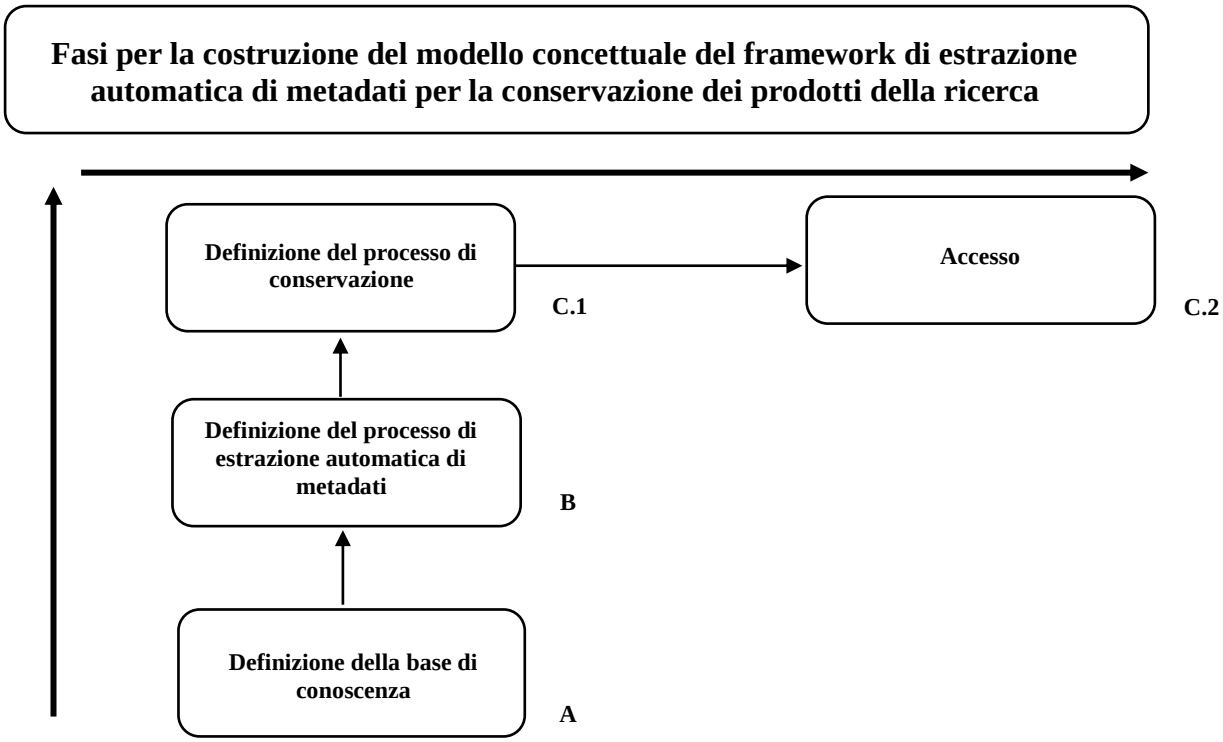

Figura 23. Le fasi che compongono il framework di estrazione automatica di metadati per la conservazione dei prodotti della ricerca

Ciascuna fase rappresenta un tassello che, relazionato agli altri, definisce l'intero framework. In tal senso, ogni fase svolge specifiche funzioni per produrre determinati output che nel complesso concorrono al raggiungimento dell'obiettivo finale.

Di seguito sono descritti le funzioni di ciascuna fase:

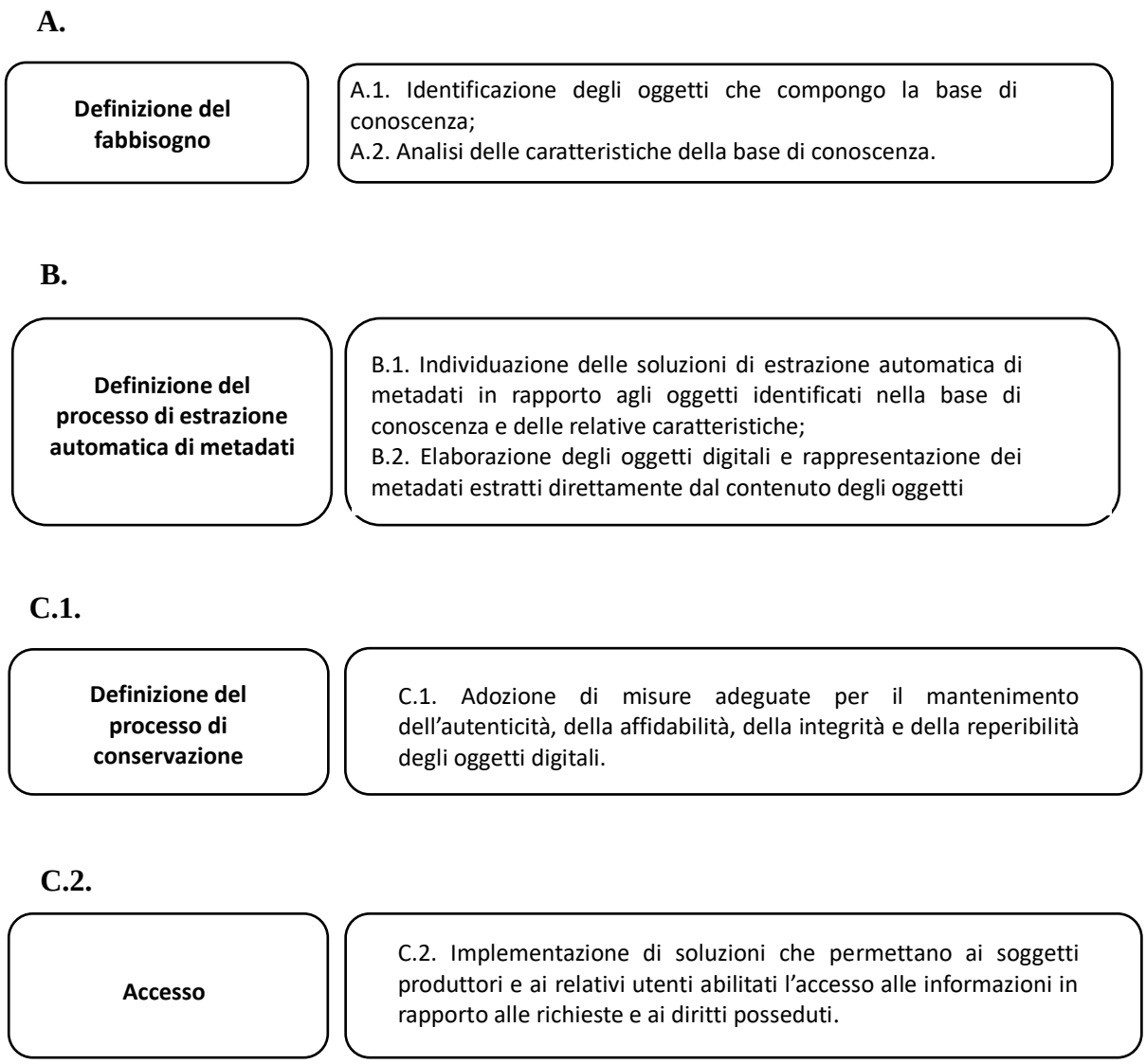

Figura 24. Fasi e obiettivi di ciascuna fase

I prossimi paragrafi intendono analizzare ciascuna fase e mettere in luce i punti cruciali da considerare nelle fasi successive.

A. Definizione del fabbisogno informativo

La definizione del fabbisogno informativo richiede due elementi essenziali relativi alla identificazione degli oggetti che compongo la base di conoscenza e all'analisi delle relative caratteristiche. Risulta pertanto essenziale quanto definito nel precedente paragrafo, il quale ha permesso prima di identificare gli oggetti digitali che rientrano nella categoria generale di prodotti della ricerca e poi di analizzarne le caratteristiche. Per esempio, tra gli oggetti rientrano i documenti. La maggior parte dei documenti afferenti ai prodotti della ricerca sono documenti nativi digitali, sebbene esistano casi in cui i documenti siano originariamente prodotti su supporto cartaceo e successivamente scannerizzati. Tale costatazione determina ricadute notevoli sul workflow e sulla scelta degli strumenti per l'estrazione automatica di metadati e ribadisce la propedeuticità delle fasi prima esposte. Inoltre, i documenti possono essere composti al loro interno da diversi oggetti informativi come, per esempio, le immagini o le tabelle. Immagini e tabelle rappresentano, a loro volta, ulteriori esempi di oggetti di conoscenza strutturata concentrata in spazi estremamente ridotti. Ignorare tali elementi determina una perdita di conoscenza significativa ed è pertanto necessario considerare nella fase soluzioni che siano in grado di identificare ed estrarre la conoscenza da tali oggetti in formati machine-readable. Altri esempi di oggetti digitali sono le basi di dati che permettono la gestione di un'ingente mole di informazioni relative agli ulteriori oggetti digitali spesso ignorate, nella convinzione che l'adozione delle tecnologie più innovative sia sufficiente a garantirne la conservazione e l'accesso nel tempo. Inoltre, alcuni degli oggetti sono prodotti su formati proprietari che potrebbero richiedere procedure di riversamento per mitigare il rischio della illeggibilità.

B. Definizione del processo di estrazione automatica di metadati

La definizione del processo di estrazione automatica di metadati richiede prima l'identificazione delle soluzioni di estrazione automatica di metadati. L'identificazione deve tenere in considerazione diversi elementi di valutazione tra cui:

1. le tipologie di prodotti della ricerca e le caratteristiche identificati nella prima fase;
2. le tipologie di metadati da estrarre ai fini della gestione del ciclo di vita degli oggetti digitali;
3. le soluzioni open-source attualmente implementate e riutilizzabili.

Tali elementi permettono di rappresentare infine i metadati estratti direttamente dal contenuto degli oggetti digitali. In particolare, l'estrazione di metadati deve

produrre un ricco set che permetta di identificare gli oggetti univocamente e in maniera persistente, che li descriva adeguatamente per garantirne il recupero e limitare i silenzi informativi, che ne assicuri la gestione del ciclo di vita fino alla conservazione e alla riproducibilità. La scelta deve considerare due requisiti fondamentali di cui devono essere in possesso le soluzioni in grado di garantire l'estrazione di tali informazioni: essere open-source ed essere riutilizzabili.

C.1. Definizione del processo di conservazione

Lo scopo di un processo di conservazione è mantenere stabili nel tempo i requisiti di autenticità, affidabilità, integrità, accessibilità e reperibilità degli oggetti digitali. Il mantenimento di tali requisiti può essere basato su diversi modelli e strategie in rapporto agli oggetti e alle relative caratteristiche. La scelta del modello e della strategia concorrono a definire il processo di conservazione.

Il modello di riferimento definito dallo standard ISO 14721:2012 OAIS[152], e imposto dal legislatore per la conservazione nel nostro Paese, costituisce indubbiamente la soluzione più funzionale grazie alla sua indipendenza dal contesto applicativo e dalle tecnologie. Tale indipendenza permette di personalizzare il modello in rapporto alle esigenze organizzative della comunità designata e di implementare le soluzioni tecnologiche più adeguate contribuendo a garantire il requisito della conservazione a lungo termine. ll processo di conservazione dei prodotti della ricerca basato sullo standard ISO 14721 OAIS prevede la generazione, durante le varie fasi del ciclo di vita degli oggetti digitali, di pacchetti informativi (SIP, AIP e DIP) composti dagli stessi oggetti più i relativi metadati estratti durante la fase precedente. Alla luce dei requisiti metodologici e tecnologici richiesti dalla realizzazione di un archivio OAIS e in considerazione dello stato dell'arte in merito alla archiviazione in repository dei prodotti della ricerca, la conservazione dovrebbe essere affidata all'esterno della struttura organizzativa.

Il processo di conservazione richiede inoltre anche l'adozione di specifiche strategie da adottare per mantenere inalterate le caratteristiche degli oggetti digitali e mitigare il rischio della obsolescenza tecnologica. Tra le strategie proposte figurano quella della emulazione e quella della migrazione. In particolare, la strategia della emulazione prevede l'implementazione di soluzioni hardware e software affinché simulino gli ambienti originariamente utilizzati per la produzione degli oggetti digitali. Il principale vantaggio offerta dalla strategia della emulazione è assicurare l'immutabilità e la riproducibilità dei contenuti e delle forme degli oggetti. Tuttavia, la complessità realizzativa legata ai molti limiti pratici che comporta l'adozione di tale strategia lascia spazio all'adozione della strategia della migrazione basata sul riversamento della tecnologia ritenuta obsoleta a quella non obsoleta mantenendo identici il contenuto e la forma. La praticità offerta dalla migrazione è indubbiamente maggiore rispetto alla

[152] In merito al modello OAIS si è parlato nel dettaglio nel paragrafo 2.2. del presente lavoro

emulazione sebbene presenti possibili rischi nella errata rappresentazione del contenuto e della forma della copia rispetto all'originale.

C.2. Accesso

L'accesso permette agli enti produttori di acquisire gli oggetti digitali conservati nell'archivio OAIS in rapporto alle richieste degli utenti abilitati. L'utente può richiedere al sistema di conservazione l'accesso agli oggetti digitali per acquisire le informazioni nei limiti previsti dalla normativa di riferimento e dalle misure tecniche-organizzative finalizzate e garantire un livello di sicurezza adeguato al rischio anche in materia di protezione dei dati personali. La richiesta di esibizione è basata sull'uso di metadati associati all'oggetto in fase di formazione. Risulta pertanto cruciale pianificare fina dalla fase di formazione degli oggetti i criteri necessari per assicurare l'accesso e il recupero in fase di conservazione. Infine, l'esibizione degli oggetti avviene tramite DIP che contiene gli oggetti digitali ed eventualmente i relativi metadati.

L'analisi delle fasi appena esposta ha condotto alla identificazione dei seguenti elementi per la costruzione del modello concettuale del framework:

1. 4 "Moduli" relativi alle funzioni concettuali realizzate dal framework ciascuno con le relative componenti;
2. 2 "Componenti" che potrebbero includere o meno eventuali elementi minimi;
3. 13 "Elementi" che caratterizzano l'unità minima di cui si compone ciascun modulo o componente;
4. 3 sistemi di acquisizione, trasmissione e smistamento degli oggetti denominati "Document Management Tools";
5. 1 "Repository" per l'archiviazione temporanea degli oggetti;
6. 1 "Archivio" per la conservazione digitale a lungo termine.

L'immagine di seguito illustra l'organizzazione ad alto livello degli elementi che compongono l'architettura del framework:

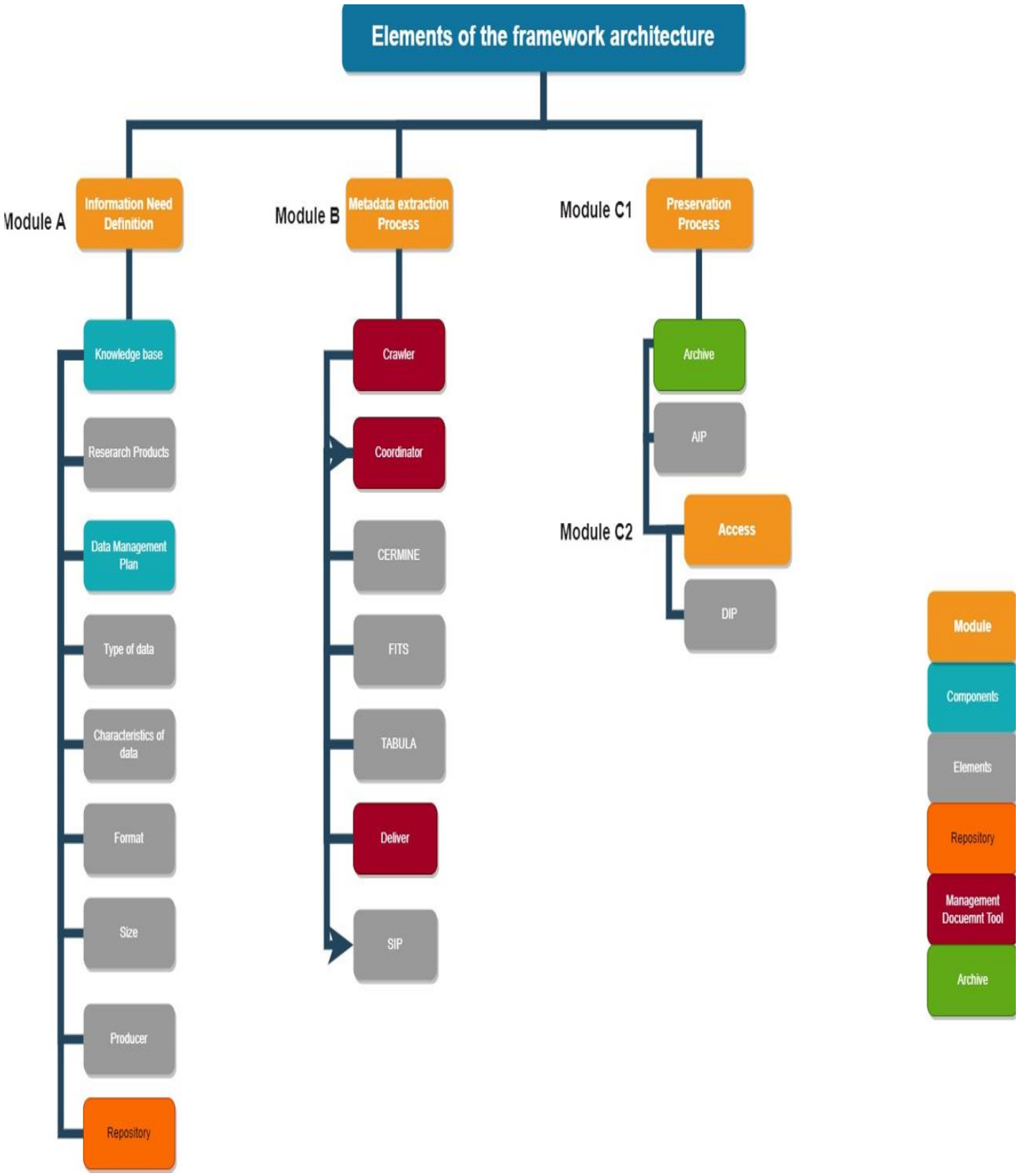

Figura 25. L'organizzazione del framework.

Tali elementi concorrono alla costruzione del modello concettuale del framework di estrazione automatica di metadati per la conservazione dei prodotti della ricerca di seguito illustrato

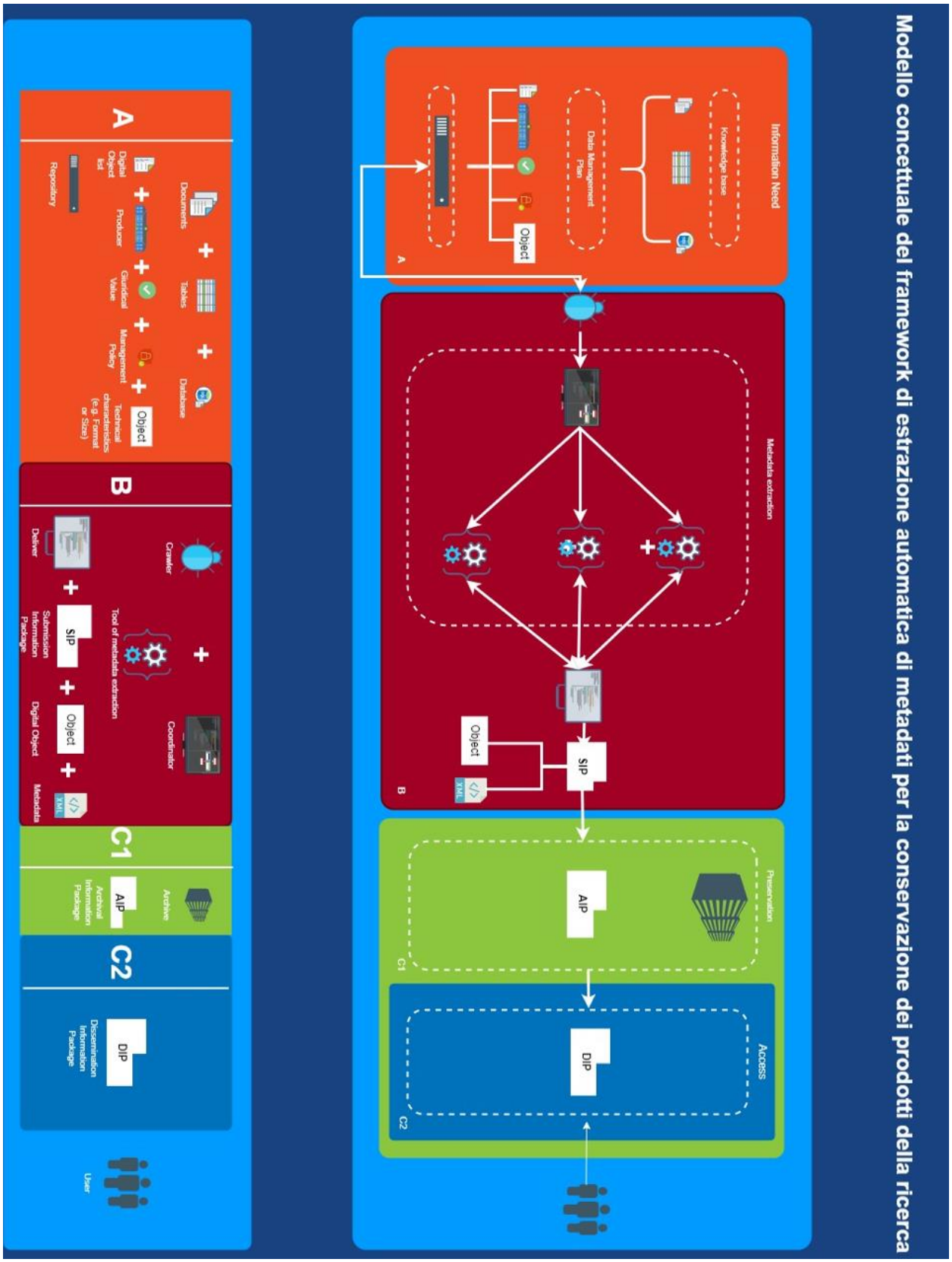

Figura 26. Il modello concettuale del framework di estrazione automatica di metadati per la conservazione dei prodotti della ricerca

2.2.2. Lo sviluppo del framework di estrazione automatica di metadati per la conservazione dei prodotti della ricerca

MEPA (Metadata Extraction, Preservation and Access) è un sistema decentralizzato per l'estrazione automatica di metadati distribuita da differenti tipologie di oggetti digitali afferenti ai prodotti della ricerca per la conservazione digitale secondo il modello nazionale definito dal legislatore. L'architettura di MEPA è incentrata su workflow modulare che permette l'eventuale personalizzazione attraverso l'integrazione di ulteriori strumenti.

In particolare, MEPA prevede quattro moduli propedeutici di seguito descritti:

Il **Modulo A** identifica la base di conoscenza e ne descrive le caratteristiche. L'identificazione della base di conoscenza dipende dalla tipologia di attività di ricerca e dalle tipologie di prodotti creati anche in rapporto all'obiettivo della ricerca stessa. La base di conoscenza di MEPA si basa su diverse tipologie di oggetti digitali tra cui documenti (che contengono altri oggetti come, per esempio, le tabelle) e le basi di dati. Tale base di conoscenza nel tempo potrà essere estesa anche ad ulteriori tipologie di prodotti della ricerca sebbene ciò renderà necessario valutarne le caratteristiche ed eventualmente integrare i tool idonei alla elaborazione. Le caratteristiche riguardano l'identificazione delle figure coinvolte nel ciclo di vita degli oggetti, la modalità di formazione, i formati, le soluzioni adottate per garantire l'efficacia probatoria, quelle per assicurare l'integrità e le misure di sicurezza. Tali caratteristiche sono formalizzate durante il modulo A nel Data Management Plan, la cui redazione è obbligatoria nel caso in cui i progetti di ricerca dovessero essere finanziati da fondi pubblici europei. L'identificazione si distingue in identificazione delle persone giuridiche e in identificazione delle persone fisiche. Le prime riguardano i soggetti produttori come, per esempio, le università o i centri di ricerca, mentre le seconde i produttori che fisicamente formano gli oggetti digitali (per esempio i ricercatori o il personale interessato dalla gestione dell'attività di ricerca che formano i documenti o che popolano le basi di dati). Tutti i documenti scelti per comporre la base di conoscenza di MEPA sono nativi digitali che possono essere formati o tramite l'utilizzo di appositi strumenti software oppure attraverso l'aggregazione di dati provenienti da una o più basi di dati. In tal senso, l'intera base di dati assume valore documentale che deve essere preservato nel tempo. La formazione degli oggetti digitali è basata poi su diversi formati tra cui **.PDF, *.JPG* e **.SQL* a cui è apposta una delle diverse firme elettroniche (avanzata, qualificata o digitale) in rapporto alla tipologia di oggetto. I documenti formati sono archiviati nel repository collocato all'interno della organizzazione del soggetto produttore che nella logica del modello della conservazione digitale coincide con il Produttore dei PdV inteso come colui che è responsabile del trasferimento degli oggetti digitali nel sistema di conservazione e che prende visione del rapporto di versamento per verificare la corretta acquisizione da parte dello stesso sistema. Il

produttore dei PdV avvia la conservazione che invoca il modulo B per l'estrazione automatica dei metadati e la loro associazione agli oggetti digitali prima di trasferirli al sistema di conservazione.

Il **modulo B** realizza l'estrazione automatica di metadati distribuita dai prodotti della ricerca attraverso 4 processi core. Il **crawler** recuperare i prodotti della ricerca depositati nel repository dell'ente produttore attraverso i sistemi di protocollo HTTP o FTP. L'acquisizione degli oggetti da parte del crawler invoca un **Coordinator** che ha la funzione di identificare la tipologia di oggetto digitale e assegnarla al sistema di estrazione automatica di metadati più idoneo. Il processo di estrazione automatica di metadati attiva uno o più sistemi che possono elaborare in modo sincrono o asincrono. In particolare, **CERMINE** è funzionale all'estrazione di metadati descrittivi, strutturali e in parte di metadati amministrativi dai documenti nativi digitali prodotti in formato PDF. **FITS** estrae un complesso set di metadati amministrativi con informazioni tecniche anche ai fini della conservazione da tutte le tipologie di oggetti digitali indipendentemente dal formato. **TABULA** permette l'estrazione automatica di metadati strutturali dalle tabelle poiché spesso ignorate o inadeguatamente estratte dai precedenti sistemi. I metadati estratti dai sistemi di estrazione automatica sono rappresentanti ciascuno in un file XML recepito da un **Deliver** che li associa all'oggetto digitale recuperato direttamente dal Coordinator. Oggetti e metadati associati formano il Pacchetto di Versamento trasferito al sistema di conservazione che da avvio al **Modulo C.1.**

Il **Modulo C.1.** svolge la funzione centrale del framework relativa al processo di conservazione degli oggetti digitali in capo al responsabile del servizio di conservazione di conservazione ed è realizzata all'interno del sistema di conservazione secondo quanto previsto dal modello nazionale. In particolare, il sistema di conservazione acquisisce il **pacchetto di versamento** trasferito dal Deliver tramite web services. La modalità via Web Service garantisce infatti l'interfacciamento diretto tra gli applicativi permettendo il versamento degli oggetti dal soggetto produttore al sistema di conservazione che esegue le verifiche sulla composizione del pacchetto di versamento, sull'integrità dei file e sull'insieme dei metadati. In particolare, le validazioni sulle verifiche riguardano:

- Validazione sulla congruità del pacchetto di versamento: la validazione del pacchetto di versamento è il risultato ottenuto al termine della verifica da parte del sistema di conservazione della congruità tra l'indice dei metadati con il numero degli oggetti che compongo il pacchetto.
- Validazione sulle specifiche del singolo oggetto digitale: la validazione delle specifiche relative al singolo oggetto digitale permette di verificare:

- La presenza del viewer e dei MIME type nel sistema di conservazione;
- La validità della firma;
- La validità della eventuale marca temporale;
- Congruità tra l'impronta calcolata e associati agli oggetti digitali dai sistemi di estrazione automatica di metadati e quella generata dal sistema di conservazione.

Inoltre, il sistema di conservazione identifica il processo di trasferimento, i soggetti e le tecnologie adoperate ad ogni operazione di versamento effettuata.

Nel caso in cui il pacchetto di versamento non superasse i **controlli** validazione, il sistema di conservazione notifica al Produttore dei PdV interno all'ente produttore la mancata conservazione degli oggetti digitali e il motivo dell'errore. Tale comunicazione avviene tramite corrispondenza elettronica all'indirizzo comunicato e il sistema permette al Produttore di apportare le necessarie modifiche per procedere con la conservazione degli oggetti digitali. Se il pacchetto di versamento supera le operazioni di verifica effettuate dal sistema questo genera il **rapporto di versamento** per ciascun pacchetto accettato. Il rapporto di versamento è utile tanto al soggetto produttore quanto al soggetto che eroga servizio di conservazione poiché attesta la corretta acquisizione degli oggetti da parte del sistema di conservazione. Tale rapporto contiene le indicazioni identificative di ogni oggetto digitali, le tempistiche del versamento e i dati dei soggetti che hanno realizzato il versamento e può essere visualizzato e scaricato dall'utente abilitato dal soggetto produttore direttamente dalla dashboard del sistema di conservazione. Il sistema di conservazione trasforma quindi il pacchetto di versamento in **pacchetto di archiviazione** che contiene gli oggetti digitali, i metadati estratti automaticamente arricchiti da ulteriori informazioni per la rappresentazione in rapporto alla tipologia di formato e l'XML dell'**indice di conservazione** basato sulla norma **UNI 11386:2020**. In tal senso il pacchetto di archiviazione, sottoscritto dal responsabile del servizio di conservazione, si compone di tre elementi: XML dell'UNI SInCRO, XML del set di metadati e oggetto digitale. Il SiNCRO supporta l'interoperabilità tecnica tra i sistemi di conservazione in caso di richiesta di trasferimento degli oggetti da un soggetto conservatore ad un altro, mentre i metadati associati agli oggetti digitali ne permettono il recupero.

Il pacchetto di archiviazione è considerato uno degli elementi estremamente statici del ciclo di vita degli oggetti digitali. In realtà, la gestione del pacchetto di archiviazione richiede il costante aggiornamento delle relazioni che si instaurano tra le differenti tipologie di oggetti digitali versati anche in momenti diversi ma archivisticamente legati. Inoltre, periodicamente il sistema di conservazione effettua l'analisi degli oggetti che hanno esaurito il tempo minimo di conservazione - secondo i tempi stabiliti dal piano di conservazione dell'ente - e che possono essere selezionati per lo **scarto** o per la conservazione permanente. Il sistema genera un elenco degli

oggetti destinati allo scarto che il responsabile del servizio di conservazione comunica al responsabile della conservazione del soggetto produttore. Il responsabile della conservazione trasmette tale elenco alla soprintendenza archivistica regionale per ottenere il nulla osta allo scarto che può essere totale o parziale. Ricevuto il nulla osta e l'elenco degli oggetti che possono essere scartati, il responsabile della conservazione trasmette tale elenco insieme al nulla osta e al provvedimento autorizzatorio allo scarto al responsabile del servizio di conservazione che procede ad eliminare definitivamente gli oggetti dal sistema di conservazione. Infine, durante l'attività di conservazione il responsabile del servizio di conservazione effettua i controlli relativi alla leggibilità e alla integrità degli oggetti conservati almeno all'anno. La verifica della leggibilità prevede selezione di un campione degli oggetti che sia rappresentativo di tutti i formati adottati per la formazione di tali oggetti. In caso di illeggibilità, MEPA indica gli oggetti che non è stato possibile rappresentare e il responsabile del servizio di conservazione, supportato anche dalle indicazioni fornite dall'Allegato 2 delle Linee Guida AgID, procede a riversare il formato obsoleto in un altro formato che sia *de iure,* aperto e interoperabile.

In tal senso, il responsabile del servizio di conservazione produce delle copie informatiche di documenti informatici e richiede al soggetto produttore l'attestazione del contenuto e della forma della copia rispetto all'originale mediante sottoscrizione. In caso di disconoscimento il soggetto produttore richiede la presenza di un pubblico ufficiale per attestare, attraverso il raffronto, la congruenza del contenuto e della forma tra i documenti. Tali passaggi dimostrano che la gestione del pacchetto di archiviazione non è un momentaneo deposito degli oggetti all'interno del sistema di conservazione ma al contrario si configura come una attività estremamente dinamica e sottoposto a processi che potrebbero anche minare il valore probatorio degli oggetti digitali.

Successivamente alla gestione del pacchetto di archiviazione gli utenti abilitati richiedono l'accesso agli oggetti conservati all'interno del sistema di conservazione. Tale richiesta avvia il modulo C.2.

Il **modulo C.2.** permette agli utenti autorizzati l'accesso diretto, anche da remoto, agli oggetti digitali conservati, attraverso la produzione da parte del sistema di conservazione dei **pacchetti di distribuzione** in formato *.ZIP***.** Ciascun pacchetto è sottoscritto dal responsabile del servizio di conservazione e può contenere uno o più pacchetti di archiviazione in rapporto alle richieste avanzate dagli utenti. Ogni pacchetto di distribuzione contiene pertanto duplicati informatici dal medesimo valore probatorio degli originali preservati all'interno del sistema di conservazione. La procedura di accesso avviene grazie alla abilitazione degli utenti indicati dal soggetto produttore, in qualità di soggetti autorizzati a richiedere la produzione dei pacchetti di distribuzione e dei relativi oggetti digitali, in rapporto ai diritti posseduti, al Conservatore attraverso il rilascio di credenziali (ID e password). Gli utenti abilitati si collegano all'interfaccia web del portale del sistema di conservazione utilizzato in MEPA

e utilizzano le credenziali rilasciate per richiede l'accesso agli oggetti digitali. L'accesso è quindi basato sull'adozione di misure tecniche e organizzative per garantire un livello di sicurezza congruo anche in rapporto alle disposizioni normative. Il sistema di conservazione, in particolare, traccia i log di accesso effettuati dagli utenti e periodicamente il responsabile del servizio di conservazione effettua le verifiche necessarie per garantire la sicurezza del sistema di conservazione. Tali verifiche riguardano, per esempio, ripetuti tentativi di accesso fallito da parte di dispositivi non identificati o possibili data breach. In tal senso, il Conservatore redige e adotta il Piano della sicurezza in cui descrive le misure adottate per assicurare l'integrità degli oggetti digitali ed evitare l'eventuale fuga di dati.

Piano di misurazione della qualità dei risultati

Il passaggio futuro riguarda la misurazione della qualità dei risultati ottenuti al termine del processo di elaborazione di MEPA. In particolare, la misurazione mira a verificare l'efficacia e l'efficienza di MEPA rispetto ai risultati ottenuti al termine del processo di estrazione automatica di metadati e di conservazione degli oggetti digitali.

Dal punto di vista metodologico, il calcolo dei risultati relativi all'estrazione automatica di metadati è basato sull'uso della:

- Precision;
- Recall.

Le percentuali ottenute sono quindi valutate in rapporto quantità di metadati estratti rispetto a quelli estraibili da ciascun sistema e alla qualità dell'estrazione che tiene conto della corretta corrispondenza tra metadato e valore informativo associate anche. Il calcolo dei risultati relativi alla corretta esecuzione del processo di conservazione è basato sulla produzione del pacchetto di distribuzione e sulla corretta creazione di tale pacchetto e quindi degli oggetti digitali scopo della richiesta di accesso da parte dell'utente.

Bibliografia

- ACKOFF, Russell Lincoln, "Ackoff From Data to Wisdom", in *Journal of Applied System Analysis*, 1989;
- Agenzia per l'Italia Digitale, "Allegato 1. Glossario dei Termini e degli Acronimi" in Linee Guida sulla formazione, gestione e conservazione dei documenti informatici, Roma, maggio 2021;
- Agenzia per l'Italia Digitale, "Allegato 2. Formati di file e riversamento" in Linee Guida sulla formazione, gestione e conservazione dei documenti informatici, Roma, 2021;
- Agenzia per l'Italia Digitale, "Allegato 3. Certificazione di processo" in Linee Guida sulla formazione, gestione e conservazione dei documenti informatici, Roma, 2021;
- Agenzia per l'Italia Digitale, "Allegato 4. *Standard e specifiche tecniche*" in Linee Guida sulla formazione, gestione e conservazione dei documenti informatici, Roma, 2021;
- Agenzia per l'Italia Digitale, "Allegato 5. Metadati" in Linee Guida sulla formazione, gestione e conservazione dei documenti informatici, Roma, 2021;
- Agenzia per l'Italia Digitale, "Allegato 6. Comunicazione tra AOO di Documenti Amministrativi Protocollati" in Linee Guida sulla formazione, gestione e conservazione dei documenti informatici, Roma, 2021;
- Agenzia per l'Italia Digitale, "Elenco dei conservatori iscritti al Marketplace dei servizi di conservazione", link: https://conservatoriqualificati.agid.gov.it/?page_id=276. Consultato il 27/02/2023;
- Agenzia per l'Italia Digitale, "FAQ", link: https://www.agid.gov.it/it/domande-frequenti/documento-informatico. Consultato il 27/02/2023;
- Agenzia per l'Italia Digitale, "Linee Guida sulla formazione, gestione e conservazione dei documenti informatici", link: https://www.agid.gov.it/sites/default/files/repository_files/linee_guida_sul_documento_informatico.pdf. Consultato il 27/02/2023;
- Agenzia per l'Italia Digitale, "Modelli di interoperabilità tra sistemi di conservazione", Roma, 2021;
- Agenzia per l'Italia Digitale, "Regolamento sui criteri per la fornitura dei servizi di conservazione dei documenti informatici", Roma, 2021;
- Agenzia per l'Italia Digitale, "Vademecum per l'implementazione delle linee guida sulla formazione, gestione e conservazione dei documenti informatici, 2022, link: *https://www.agid.gov.it/it/linee-guida*. Consultato il 27/02/2023;
- ALTO, Palo, "Layout and Content Extraction for PDF Documents", in *Verlag Berling Heidelberg*, *Springer*, 2004. pp. 213-224;
- Archival Portal Europe, "The history of Europe - one click away", link: https://www.archivesportaleurope.net/. Consultato il 27/02/2023;

- Archival Portal Europe, “The use of EAG in Archives Portal Europe”, link: https://www.archivesportaleurope.net/tools/for-content-providers/standards/eag/, Consultato il 27/02/2023;
- Archivio Federale Svizzero, “SIARD Suite”, link: https://www.bar.admin.ch/bar/it/home/archiviazione/strumenti/siard-suite.html. Consultato il 27/02/2023;
- BACA, Murtha, “Introduction to Metadata. Second Edition”, in *The Getty Research Institute*, (CA 90049-1682), 2008. p.7-19;
- BINMAKHASHEN, Galal M., MAHMOUD, Sabri A., “Document Layout Analysis: A comprehensive Survey”, in *ACM Computing Surveys,* Vol. 52, No. 6, 2019;
- BRAINARD, Jeffrey, “Scientists are drowning in COVID-19 papers. Can new tools keep them afloat? The hunt is on for better ways to collect and search pandemic studies”, in *Science*, 2020, link : https://www.science.org/content/article/scientists-are-drowning-covid-19-papers-can-new-tools-keep-them-afloat . Consultato il 27/02/2023;
- CARUCCI, Paola, GUERCIO, Maria, *Manuale di Archivistica. Nuova Edizione*, Roma, Carocci Editore, 2021, p. 384;
- CASSELLA, Maria, *Open Access e comunicazione scientifica*, Editrice Bibliografica, Milano, 2012;
- CHRISTOPHIDES, Vassilis, BUNEMAN, Peter, “Report on the First International Workshop on Database Preservation (PresDB’07)”, in *SIGMOD Record*, Vol. 36, No. 3, 2007;
- Commissione Europea, “Comunicazione della commissione al parlamento europeo, al consiglio, al comitato economico e sociale europeo e al comitato delle regioni. Un'agenda digitale europea”, Bruxelles, 19 maggio 2010;
- Commissione Europea, “CORDIS”, link: https://cordis.europa.eu/about/it. Consultato il 27/02/2023
- Commissione Europea, “Decennio Digitale Europeo: Obiettivi digitali per il 2030”, link: https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/europes-digital-decade-digital-targets-2030_it. Consultato il 27/02/2023;
- Commissione Europea, “Digital Europe Programme”, link: https://ec.europa.eu/info/funding-tenders/find-funding/eu-funding-programmes/digital-europe-programme_it. Consultato il 27/02/2023;
- Commissione Europea, “FOSTER - Facilitate Open Science Training for Euorpean Research”, link: https://cordis.europa.eu/project/id/612425. Consultato il 27/02/2023;
- Commissione Europea, “Horizon 2020”, link: https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-2020_en. Consultato il 27/02/2023;
- Commissione Europea, “i2010: Digital Libraries”, link: https://eur-lex.europa.eu/EN/legal-content/summary/i2010-digital-libraries.html. Consultato il 27/02/2023;

- Commissione Europea, “Project Databases”, link: https://research-and-innovation.ec.europa.eu/projects/project-databases_en. Consultato il 27/02/2023;
- Commissione Europea, “Validation of the results of the public consultation on Science 2.0: Science in Transition”, link: https://www.espci.psl.eu/sites/www.espci.psl.eu/IMG/pdf/science_2_0_final_report.pdf. Consultato il 27/02/2023;
- COUNCILL, Isaac, GILES, Lee C., KAN, Min-Yen, “ParsCit”, link: https://github.com/knmnyn/ParsCit. Consultato il 27/02/2023
- FORCE11, “The Fair Data Principles”, link: https://www.force11.org/group/fairgroup/fairprinciples. Consultato il 27/02/2023;
- Consiglio dell’Unione Europea, “Proposta di decisione del consiglio relativa all'istituzione del programma specifico di attuazione di Orizzonte Europa - il programma quadro di ricerca e innovazione”, link: https://data.consilium.europa.eu/doc/document/ST-8550-2019-INIT/it/pdf. Consultato il 27/02/2023;
- CUONG, Nguyen Viet, CHANDRASEKARAN, Muthu Kumar, KAN, Min-Yen, LEE, Wee Sun, “Scholarly Document Information Extraction using Extensible Features for Efficient Higher Order Semi-CRFs*”, in JCDL’15, ACM*, 2015. DOI: *http://dx.doi.org/10.1145/2756406.2756946*;
- Digital Bibliography & Library Project, “Statistics – Records in DBLP”, link: https://dblp.uni-trier.de/statistics/recordsindblp.html. Consultato il 27/02/2023;
- Dublin Core, “Dublin Core”, link: https://www.dublincore.org/. Consultato il 27/02/2023;
- GIGLIA, Elena, “Open Science dalla A alla Z”, in ZENODO, 2020, link: https://zenodo.org/record/3907297#.Y9aqf3CZND9 . Consultato il 27/02/2023;
- ELSHAVWI, Radwa, SAKR, Sherif, TALIA, Domenico, TRUNFIO, Paolo, “Big Data Systems Meet Machine Learning Challenges: Towards Big Data Science as a Service”, in *Journal of Big Data Research, Elsevier*, 2018;
- Ente Italiano di Normazione, “UNI Home”, link: https://www.uni.com/index.php Consultato il 27/02/2023;
- ESKENAZI, Sébastien, KRÄMER, Petra Gomez, OGIER, Jean-Marc, “A comprehensive survey of mostly textual document segmentation algorithms since 2008”, *in Pattern Recognition, Elsevier*, 2016. DOI: http://dx.doi.org/10.1016/j.patcog.2016.10.023;
- GARTNER, Richard, “Metadata – Shaping Knowledge from Antiquity to the Semantic Web”, in *Springer*, 2016. DOI:10.1007/978-3-319-40893-4;
- GINSPARG, Paul, “arXiv”, 1991, link: arXiv.org. Consultato il 27/02/2023;
- GOLUB, Koraljka, MULLER, Henk, TONKIN, Emma, “Technologies for Metadata Extraction”, in *Semantics and Ontologies*, *World Scientific,* 2014. ISBN: 9789812836298;

- GREENBERG, Jane, “Big Metadata, Smart Metadata, and Metadata Capital: Toward Greater Synergy Between Data Science and Metadata”, *Journal of Data and Information Science*, 2017;
- GUERRINI, Mauro, *Gli Archivi Istituzionali*, Editrice Bibliografica, Milano, 2010;
- HAN, Han, GILES, C.Lee, MANAVOGLU, Eren, ZHA, Hongyuan, ZHANG, Zhenyue, FOX, Edward A., “Automatic document metadata extraction using support vector machines”, in *Joint Conference on Digital Libraries, organized by ACM/IEEE*, 2033. pp. 37–48;
- a cura di HESS, Charlotte, OSTROM., Elinor, *La conoscenza come bene comune. Dalla teoria alla pratica*, Bruno Mondadori, Milano, 2009. ISBN: 978-88-6159-142-4;
- IEEE, “Standards (IEE BIG DATA)”, link: https://bigdata.ieee.org/standards, Consultato il 27/02/2023;
- ISTI-CNR, “Welcome to the DELOS Network of Excellence”, link: http://delosw.isti.cnr.it/. Consultato il 27/02/2023;
- Istituto Centrale per il Catalogo Unico, “ICCU Home”, link: https://www.iccu.sbn.it/it/. Consultato il 27/02/2023;
- International Standardization for Standardization, “ISO”, link: https://www.iso.org/home.html. Consultato il 27/02/2023;
- ISO 23081-1/2:2006-2007 Information and documentation - Records management - Metadata for records, Part. 1: Principles - Part. 2: Conceptual and implementation issues;
- ISO 32000-1:2008 Document management — Portable document format — Part 1: PDF 1.7;
- ISO 14721:2012 Space data and information transfer systems — Open archival information system (OAIS) — Reference model;
- ISO 15489-1:2016 Information and Documentation – Records Management – Part:1 Concepts and principle;
- LANDINO, Costantino, MARZOTTI, A. Pasqualina, *Memorie dinamiche. La conservazione dei database e il web archiving*, Edizioni ANAI, Roma, 2018;
- Library of Congress, “Library of Congress Home”, link: https://www.loc.gov/. Consultato il 27/02/2023;
- Library of Congress Network Development and MARC Standards Office, “Capire PREMIS”, link: https://www.loc.gov/standards/premis/Understanding-PREMIS_italian.pdf. Consultato il 27/02/2023;
- Library of Congress Network Development and MARC Standards Office, “Metadata encoding and transmission standard: primer and reference manual”, link: https://www.loc.gov/standards/mets/METSPrimer.pdf. Consultato il 27/02/2023;
- Library of Congress Network Development and MARC Standards Office, “METS. Metadata Encoding & Trasmission Standard Official Web site” link: https://www.loc.gov/standards/mets/. Consultato il 27/02/2023;

- LOPEZ, Patrice, "GROBID", link: https://grobid.readthedocs.io/en/latest/Introduction/ Consultato il 27/02/2023;
- MATHIS, Christian, "Data Lakes", in *Springer*, 2017;
- METITIERI, Fabio, RIDI, Riccardo, *Biblioteche in rete. Istruzioni per l'uso*, Laterza, 2007. EAN: 9788842076636;
- MICHETTI, Giovanni, "Il modello OAIS", *DigItalia,* Anno III, Numero I, 2008. p.35;
- Ministère de L'Enseignement supérieur et de la Recherche, "Second National Plan for Open Science", link: https://www.ouvrirlascience.fr/second-national-plan-for-open-science-2021-2024/. Consultato il 27/02/2023;
- Ministero dell'Università e della Ricerca, "Piano Nazionale per la Scienza Aperta", link: https://www.mur.gov.it/sites/default/files/2022-06/Piano_Nazionale_per_la_Scienza_Aperta.pdf. Consultato il 27/02/2023;
- Ministry of Education and Science of Ukraine, "Ukraine has joined the EU countries that have an approved plan for implementing the Open Science principles", link: https://www.kmu.gov.ua/en/news/ukraina-pryiednalas-do-krain-ies-shcho-maiut-zatverdzhenyi-plan-realizatsii-pryntsypiv-vidkrytoi-nauky. Consultato il 27/02/2023;
- NAMBOODIRI, Anoop M., JAIN, Anil K., "Document Structure and Layout Analysis", in ResearchGate, 2007. DOI: 10.1007/978-J-84628-726-8_2;
- National Library of Medicine, "About PMC", 2000, link: https://www.ncbi.nlm.nih.gov/pmc/about/intro/. Consultato il 27/02/2023;
- Online Computer Library Center, "RLG Best practice Guidelines for Encoded Archival Description", link: https://www.oclc.org/content/dam/research/activities/ead/bpg.pdf. Consultato il 27/02/2023;
- Online Computer Library Center, "Online Computer Library Center Home", link: https://www.oclc.org/en/home.html . Consultato il 27/02/2023;
- OpenAIRE, "ZENODO", link: https://zenodo.org/. Consultato il 27/02/2023;
- Open Archive Initiative, "Standard for Web Content Interoperability", link: www.openarchives.org. Consultato il 27/02/2023;
- OpenScience.eu, "Open Innovation Open Science Open to the World. A vision For Europe", link: https://openscience.eu/open-science-policy-platform-final-report. Consultato il 27/02/2023;
- Parlamento Europeo, "Regolamento (UE) n. 910/2014 del Parlamento europeo e del consiglio del 23 luglio 2014 in materia di identificazione elettronica e servizi fiduciari per le transazioni elettroniche nel mercato interno e che abroga la direttiva 1999/93/ce", Bruxelles, 28 agosto 2014;
- Parlamento Europeo, "Regolamento (UE) n. 2016/679 del Parlamento europeo e del consiglio del 27 aprile 2016 relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati e che

abroga la direttiva 95/46/CE (regolamento generale sulla protezione dei dati)*",* Bruxelles, 27 aprile 2016*;*

- Parlamento Europeo, "Regolamento (UE) n. 2018/1807 sulla libera circolazione dei dati non personali che consente alle imprese e alle amministrazioni pubbliche di archiviare e trattare i dati non personali ovunque scelgano di farlo nell'UE*"*, Bruxelles, 14 novembre 2018;
- Parlamento Europeo, "Regolamento sulla cybersicurezza (UE) n. 881/2019, che rafforza l'Agenzia dell'Unione europea per la cybersicurezza (ENISA) e istituisce un quadro per la certificazione della cybersicurezza di prodotti e servizi*"*, Bruxelles, 17 aprile 2019;
- Parlamento Europeo, "Proposta di Direttiva del Parlamento Europeo e del Consiglio relativa al riutilizzo dell'informazione del settore pubblico", Bruxelles, April 25, 2018;
- Parlamento Europeo, "Direttiva sull'apertura dei dati (UE) n. 2019/2014 che stabilisce norme comuni per un mercato europeo per i dati in possesso del governo", Bruxelles, 20 giugno 2019;
- Parlamento italiano, "D. lgs. recante il Codice dei beni culturali e del paesaggio, ai sensi dell'articolo 10 della legge 6 luglio 2002, n. 137", 22 gennaio 2004, n. 42. (in G. U. n. 45 del 24 febbraio 2004 – Suppl. Ordinario n. 28);
- Parlamento italiano, "D. lgs. recante il Codice dell'Amministrazione Digitale", 7 marzo 2005, n. 82 (in G.U. n.112 del 16-05-2005 - Suppl. Ordinario n. 93);
- PARK Jung-ran, BRENZA, Andrew, "Evaluation of Semi---Automatic Metadata Generation Tools: A Survey of The Current State of The Art*",* in *Information technology and libraries*, 2015. DOI: 10.6017/ital.v34i3.5889;
- PIGLIAPOCO, Stefano, "Il Modulo Access del Modello OAIS", in *AIDAInformazioni*, 2016. ISBN: 978-88-548-9402-0;
- PIGLIAPOCO, Stefano, "La Conservazione Digitale in Italia. Riflessioni su modelli, criteri e soluzioni", in *JLis.it,* 2019. DOI: 10.4403/jlis.it-12521;
- Presidente del Consiglio dei Ministri, "Regole tecniche per la generazione, apposizione e verifica delle firme elettroniche avanzate, qualificate e digitali", 22 febbraio 2013 (in Gazzetta Ufficiale il 21-5-2013 n. 117);
- Presidente del Consiglio dei Ministri, "Regole tecniche in materia di sistema di conservazione", 3 dicembre 2013 (in Gazzetta Ufficiale il 12-3-2014 n. 59);
- Presidente del Consiglio dei Ministri, "Regole tecniche per il protocollo informatico", 3 dicembre 2013 (in Gazzetta Ufficiale il 12-3-2014 n.59);
- Presidente del Consiglio dei Ministri, "Regole tecniche in materia di formazione, trasmissione, copia, duplicazione, riproduzione e validazione temporale dei documenti informatici nonché di formazione e conservazione dei documenti informatici delle pubbliche amministrazioni", 13 novembre 2014 (in Gazzetta Ufficiale il 12-01-2015 n. 8);

- Presidente della Repubblica, “Decreto del Presidente della Repubblica recante il Testo unico delle disposizioni legislative e regolamentari in materia di documentazione amministrativa”, 28 dicembre 2000, n. 445 (G.U. n. 42 del 20 febbraio 2001, s.o. 30/L);
- Presidente della Repubblica, “Decreto del Presidente della Repubblica recante il Regolamento recante disposizioni per l'utilizzo della posta elettronica certificata a norma dell'articolo 27 della legge 16 gennaio 2003, n. 3”, 11 febbraio 2005, n. 68 (G.U. n.97 del 28-04-2005);
- Publications Office of the European Union, “The Netherlands’ plan on open science. Open science monitor case study”, link: https://op.europa.eu/en/publication-detail/-/publication/20d4026e-4478-11e9-a8ed-01aa75ed71a1/language-en. Consultato il 27/02/2023;
- RAMALHO, Jose Carlos, FERREIRA, Miguel, FARIA, Luis, CASTRO, Rui, “Relational Database Preservation Through XML Modelling”, in *Extreme Markup Languages 2007*, Montréal, 2007;
- RAMALHO, Jose Carlos, FARIA, Luis, SILVA, Hélder, Coutada, Miguel, "DB-Preservation Toolkit in GitHub), http://keeps.github.io/db-preservation-toolkit/ attualmente migrato su: https://github.com/keeps/dbptk-developer. Consultato il 27/02/2023
- Research Libraries Group, “RLG Home”, link: https://www.rlg.org/. Consultato il 27/02/2023;
- ROVELLA, Anna, “La posta elettronica negli archivi di persona: conservazione e accesso”, AIDA Informazioni, 2022. p.p. 113-128. DOI: 10.57574/596516316;
- QUIX, Christoph, HAI, Rihan, VATOV, Ivan, “Metadata Extraction and Management in Data Lakes With GEMMS”, in *Complex Systems Informatics and Modeling Quarterly*, 2017. Pages 67–83;
- SKLUZACEK, Tyler J., “Dredging a Data Lake: Decentralized Metadata Extraction”, in *Middleware Doctoral Symposium*, 2019;
- SKLUZACEK, Yler J., KUMAR, Rohan, CHARD, Ryan, HARRISON, Galen, BECKMAN, Paul, CHARD, Kyle, FOSTER, Ian T, “Skluma: An extensible metadata extraction pipeline for disorganized data”, in *IEEE 14th International Conference on e-Science*, 2018;
- Society of American Archivists, “Society of American Archivists Home”, link: https://www2.archivists.org/. Consultato il 27/02/2023;
- The Library of Congress, “<ead>. Encoded Archival Description”, link: https://www.loc.gov/ead/. Consultato il 27/02/2023;
- The Library of Congress, “METS. Metadata Encoding & Trasmission Standard Official Web site”, link: https://www.loc.gov/standards/mets/. Consultato il 27/02/2023;
- The Library of Congress, “PREMIS. Preservation Metadata Maintenance Activity”, link: https://www.loc.gov/standards/premis/. Consultato il 27/02/2023;
- TKACZYK, Dominik, SZOSTEK, Paweł, FEDORYSZAK, Mateusz, DENDEK, Piotr Jan, BOLIKOWSKI, Łukasz, “CERMINE: automatic extraction of structured metadata from

scientific literature", in *International Journal on Document Analysis and Recognition*, *Springer*, 2015;

- UNESCO, "Open Science. UNESCO Recommendation on Open Science", link: https://en.unesco.org/science-sustainable-future/open-science/recommendation. Consultato il 27/02/2023;
- UNI, "Supporto all'Interoperabilità nella Conservazione e nel Recupero degli Oggetti digitali. SInCRO", N. 11386:2020, 2020;
- Unione Europea, "OpenAIRE", link: https://www.openaire.eu/. Consultato il 27/02/2023;
- Università di Bologna, Brescia, Calabria, Firenze, Foggia, Genova, Insubria, Lecce, Messina, Milano, Milano Bicocca, Milano Politecnico, Milano Vita-Salute San Raffaele, Modena, Molise, Napoli Federico II, Napoli L'Orientale, Napoli Partenope, Padova, Palermo, Parma, Piemonte Orientale, Roma LUMSA, Roma Tor Vergata, Roma III, Siena, Torino, Trieste, Trieste SISSA, Tuscia, Venezia IUAV, e l'Istituto Italiano di Medicina Sociale di Roma, "Dichiarazione di Messina. Documento Italiano a Sostegno della Dichiarazione di Berlino sull'Accesso Aperto alla Letteratura Accademica", 2004. Link: http://www.sssup.it/UploadDocs/7109_Dichiarazione_di_Messina.pdf. Consultato il 27/02/2023;
- Università di Harvard, "File Information Tool Set (FITS). Documentation and official code releases of the FITS and the FITS Web Service projects (File Information Tool Set (FITS). Documentation and official code releases of the FITS and the FITS Web Service projects", link: https://projects.iq.harvard.edu/fits/home . Consultato il 27/02/2023;
- Università di Porto, "RODA project", link: http://www.roda-community.org. Consultato il 27/02/2023;
- VITALI, Stefano, *Passato digitale*, Bruno Mondadori, Milano, 2004. p. 189;
- W3Schools, "HTML", link: https://www.w3schools.com/html/. Consultato il 27/02/2023;
- Wikipedia, "Encoded Archival Description", link: https://en.wikipedia.org/wiki/Encoded_Archival_Description#History. Consultato il 27/02/2023;
- WILKINSON, Mark D., DUMONTIER, Michel, AALBERSBERG, IJsbrand, APPLETON, Jan Gabrielle, AXTON, Myles, BAAK, Arie, BLOMBERG, Niklas, BOITEN, Jan-Willem, DA SILVA SANTOS, Luiz Bonino, BOURNE, Philip E, BOUWMAN, Jildau, BROOKES, Anthony J, CLARK, Tim, CROSAS, Mercè, DILLO, Ingrid, DUMON, Olivier, EDMUNDS, Scott, EVELO, Chris T, FINKERS, Richard, GONZALEZ-BELTRAN, Alejandra, GRAY, Alasdair J.G., GROTH, Paul, GOBLE, Carole, GRETHE, Jeffrey S, HERINGA, Jaap, HOEN, Peter A.C 't, HOOFT, Rob., KUHN, Tobias, KOK, Ruben, KOK, Joost, LUSHER, Scott J., MARTONE, Maryann E., MONS, Albert, PACKER, Abel L., PERSSON, Bengt, ROCCA-SERRA, Philippe, ROOS, Marco, SCHAIK, Rene van, SANSONE, Susanna-Assunta, SCHULTES ,Erik, SENGSTAG, Thierry, SLATER, Ted., STRAWN, George, SWERTZ, Morris, THOMPSON,

Mark, VAN DER LEI, Johan, VAN MULLIGEN, Erik, VELTEROP, Jan, WAAGMEESTER, Andra, WITTENBURG, Peter, WOLSTENCROFT, Katherine, ZHAO, Jun, MONS, Barend, “The FAIR Guiding Principles for scientific data management and stewardship”, in *Scientific data*, Nature. DOI: 10.1038/sdata.2016.18.