ICAR
ISTITUTO CENTRALE PER GLI ARCHIVI

ReCAP
Rete per la Conservazione e l'Accesso ai Patrimoni digitali

Conservazione dei database e Web Archiving

Costantino Landino

Archivio Centrale dello Stato
Roma, 11 Aprile 2017

ICAR
ISTITUTO CENTRALE PER GLI ARCHIVI

Agenda

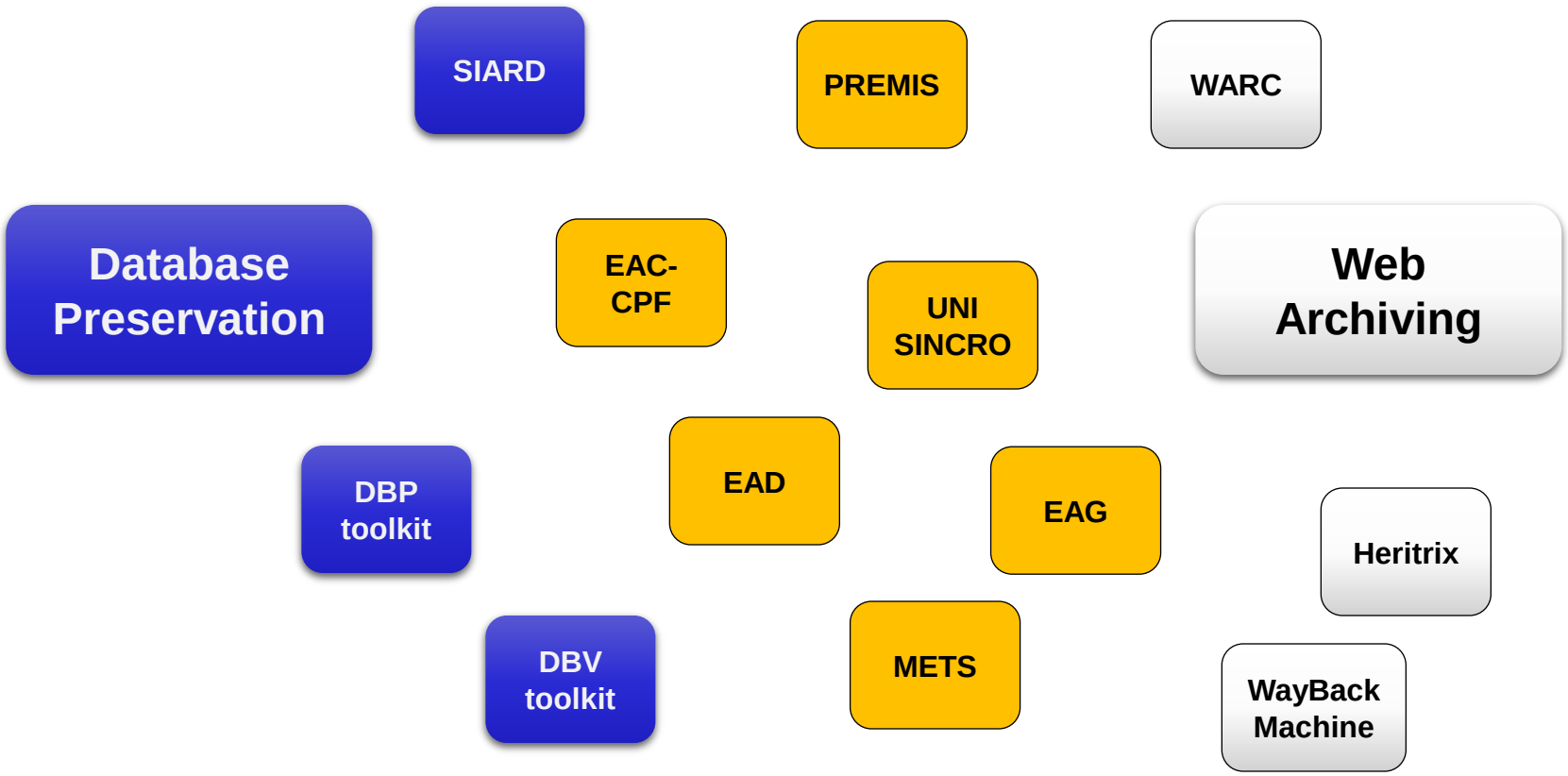
- ✓ Premesse, metadati di conservazione OAIS, PREMIS, UNISINCRO, formati di memorizzazione, Documenti informatici
- ✓ Conservazione dei database -- metadati descrittivi e di conservazione
- ✓ Strumenti software:: SIARD, Database Preservation Toolkit e Database Preservation Toolkit
- ✓ Conservazione dei contenuti di un database
- ✓ Software: RODA-IN e produzione di SIP E-Ark
- ✓ Considerazioni

ICAR
ISTITUTO CENTRALE PER GLI ARCHIVI

Agenda

- ✓ Web Archiving
- ✓ Conservazione dei siti web
- ✓ Formato WARC
- ✓ Strumenti Heritrix e Wayback machine
- ✓ Integrazione con il sito dell'ICAR e considerazioni finali

Keywords

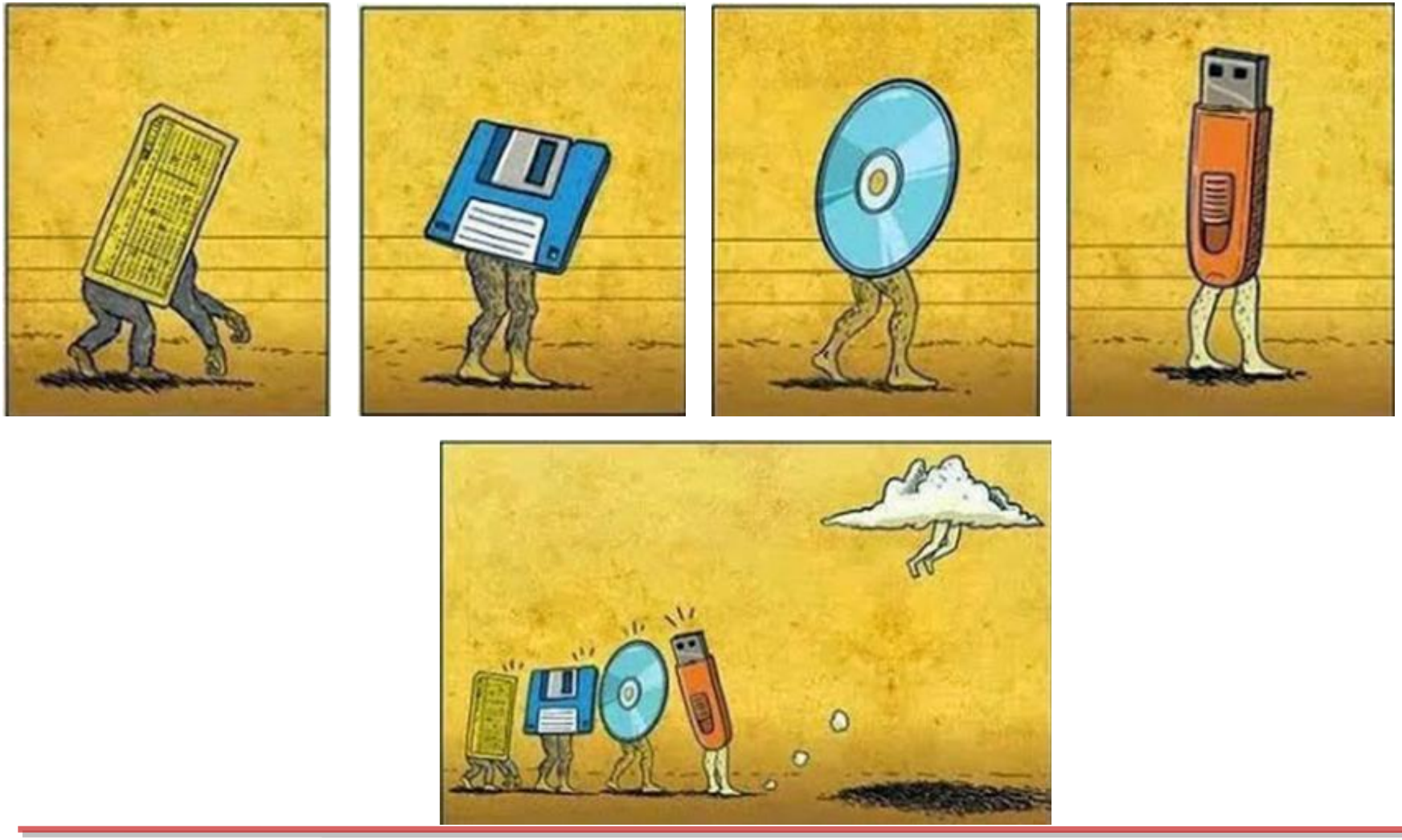

SIARD
PREMIS
WARC
Database Preservation
EAC-CPF
UNI SINCRO
Web Archiving
DBP toolkit
EAD
EAG
Heritrix
DBV toolkit
METS
WayBack Machine

Conservazione dei contenuti digitali: esplosione

Conservazione dei contenuti digitali: intelligibilità

- E’ un Libro. Si può leggerlo senza bisogno di uno schermo.
- Le pagine sono tutte accessibili e visibili. Non scompaiono in caso di mancanza di corrente.
- E’ più leggero di un portatile e non sarà obsoleto il prossimo mese.
- Può anche prestarlo a suo padre senza dovergli spiegare come usarlo.

Conservazione dei contenuti digitali: obsolescenza



Conservazione dei contenuti digitali: perdita

Conservazione dei contenuti digitali: lavoro

Conservazione database

Database

Conservazione contenuti

Conservazione web

Primo scenario

Database centrale del Sistema informativo degli Archivi di Stato (SIAS)

Conservazione del database di un sistema informativo archivistico che ha terminato il proprio ciclo di sviluppo, con un software di gestione obsoleto senza manutenzione correttiva e evolutiva da anni.

Problematiche

- Come conservare i contenuti della banca dati del sistema? Con quale formato SQL nativo, SQL standard, XML, testo?
- Cosa si perde nella trasformazione ? Quale è il livello di copertura informativa rispetto alla base dati originale?
- Quali metadati per la conservazione e la successiva reperibilità dei contenuti?
- Come costruire il pacchetto di conservazione ?

Secondo scenario

Archivio Storico Multimediale del Mediterraneo

Conservazione dei contenuti di un sistema informativo archivistico ormai su hardware e software obsoleto in fase di dismissione e senza manutenzione

Problematiche

- Estrazione dei dati in formati standard EAD, EAC-CPF, EAG ?
- Quale è il livello di copertura informativa dell'estrazione e cosa perdo nella trasformazione ?
- Come mantengo il contesto archivistico delle entità estratte ?
- Quali metadati per la conservazione e la successiva reperibilità dei contenuti?
- Come costruire il pacchetto di conservazione ?

Terzo scenario

Mostre MOVIO dell'Istituto Centrale per gli Archivi

Conservazione dei contenuti di una mostra virtuale relativa a materiale archivistico ospitata su un sistema in hosting .

Problematiche

- Come conservare i contenuti con un processo di web crawling?
- Cosa posso perdere nel recupero dei contenuti?
- Quali formati per la conservazione dei siti web?
- Quali metadati per la conservazione e la successiva reperibilità dei contenuti?
- Come costruire il pacchetto di conservazione?

Modello OAIS

OAIS (acronimo di Open Archival Information System) è lo standard ISO:14721:2003 che definisce concetti, modelli e funzionalità inerenti agli archivi digitali e gli aspetti conservazione digitale.

Il modello OAIS definisce il **pacchetto informativo** come l'entità fondamentale attorno alla quale ruotano i processi di conservazione e che è corrisponde ad un insieme logico composto dall'oggetto digitale da conservare e dai metadati necessari a garantirne la conservazione e l'accesso a lungo termine

Il pacchetto informativo risulta composto da due componenti:

- le informazioni sul contenuto (content information)
- le informazioni sulla conservazione (preservation description information). .

Modello OAIS

Le informazioni sul contenuto sono scomposte in informazioni sui dati (data object) e delle relative informazioni sulla rappresentazione (representation information) che ne permettono la comprensione.

00000000	FFFF	FFFF	FFFF	FFFF	FFFF	FFFF	FFFF	FFFF	.....
00000010	AA99	5566	31E1	1FFF	3261	0044	3281	6B00	..Uf1...2a.D2.k.
00000020	32A1	0044	32C1	6B00	32E1	0000	30A1	0000	2..D2.k.2...0...
00000030	3301	3100	3201	005F	30A1	000E	2000	2000	3.1.2.._0... . .
00000040	2000	2000	FFFF	FFFF	FFFF	FFFF	FFFF	FFFF	. .....
00000050	FFFF	FFFF	AA99	5566	30A1	0007	2000	31A1	......Uf0... .1.
00000060	0960	3141	3D08	3161	89EE	31C2	0403	D093	.`1A=.1a..1.....
00000070	30E1	00CF	30C1	0081	2000	2000	2000	2000	0...0... . . . .
00000080	2000	2000	2000	2000	2000	2000	2000	2000	. . . . .
00000090	2000	2000	2000	2000	2000	3381	3C64	3181	. . . . .3.<d1.
000000A0	0881	3421	0000	3201	001F	31E1	1FFF	3321	..4!..2...1...3!
000000B0	0005	3341	0004	3301	3100	3261	0000	3281	..3A..3.1.2a..2.
000000C0	0000	32A1	0000	32C1	0000	32E1	0000	33A1	..2...2...2...3.
000000D0	1BE2	33C2	0000	0000	2000	2000	3022	0000	..3..... . .0"..
000000E0	0000	30A1	0001	5060	0020	3209	0000	0000	..0...P`. 2.....
000000F0	0000	0000	0000	0000	0000	0000	0000	0000	.....
00000100	0000	0000	0000	0000	0000	0000	0000	0000	.....
00000110	0000	0000	0000	0000	0000	0000	0000	0000	.....

representation information

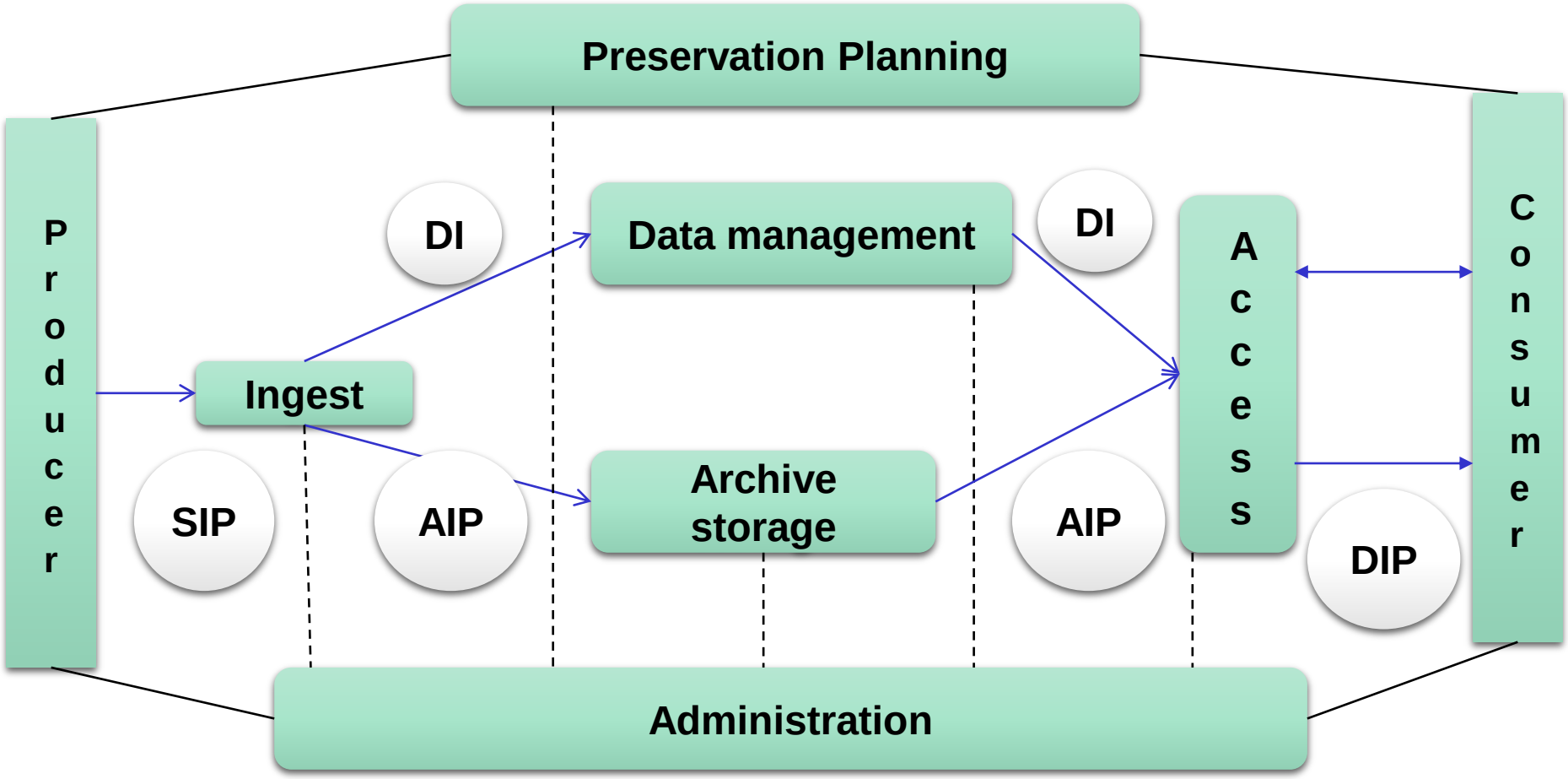
Modello OAIS

Le informazioni sulla conservazione sono finalizzate ad assicurare l'integrità delle unità documentarie singole, dei gruppi, delle relazioni di contesto e delle informazioni per l'accesso; assicurare il mantenimento nel lungo periodo in forme stabili (formati) delle modalità originarie di reperimento dei documenti e della loro accessibilità

I metadati sono raggruppati in :

- **reference information**: informazioni identificative del contenuto
- **context information:** informazioni di contesto che documentano le relazioni tra il contenuto e l'ambiente di produzione
- **provenance information**: informazioni di provenienza che documentano la storia del contenuto e le trasformazioni subite
- **fixity information**: informazioni di validazione e integrità

Modello OAIS



Modello OAIS

Gli oggetti della conservazione sono organizzati in tre tipologie di **pacchetti informativi**:

- **Pacchetti di versamento** (**SIP:** Submission Information Package): sono i pacchetti informativi inviati ad un sistema da un produttore e gestiti in fase di acquisizione. Il formato ed il contenuto possono variare in funzione delle necessità.

- **Pacchetti di distribuzione** (**DIP**, Dissemination Information Package)**:** sono i pacchetti informativi consegnati agli utenti dietro una richiesta di accesso ai contenuti. Possono essere singoli o gruppi di pacchetti.

Modello OAIS

- **Pacchetti di archiviazione (**AIP, Archival Information Package): sono il risultato della trasformazione dei pacchetti di versamento, in quanto vengono dotati di un set completo di metadati di conservazione per permettere la permanenza a lungo termine nel sistema. Un singolo pacchetto di archiviazione può contenere anche una raccolta di diversi pacchetti di versamento oppure è possibile che un singolo pacchetto di versamento debba essere frammentato in più pacchetti di archiviazione.

PREMIS

I Metadati PREMIS identificano quei metadati di conservazione necessari per il p*rocesso di conservazione digitale*, ovvero le informazioni necessarie a garantire la possibilità della tenuta, l'accessibilità, l'intelligibilità, l'autenticità delle risorse digitali. identificano **cinque aree rilevanti per la conservazione**:

- **Provenienza**: le informazioni storiche sulla custodia dell'oggetto digitale, dalla sua creazione, ogni successivo cambio di custodia fisica e/o di proprietà.

- **Autenticità**: le informazioni sufficienti a validare che l'oggetto digitale dell'archivio è proprio quello che si presuppone sia e che non sia stato alterato, intenzionalmente e non, in modo non documentato.

PREMIS

- **Attività di conservazione**: le azioni intraprese per conservare l'oggetto digitale e qualsiasi conseguenza di tali azioni che impattino su forma, percezione o funzionalità dell'oggetto.

- **Ambiente tecnologico**: hardware, sistema operativo e applicazioni software necessarie a rappresentare e usare l'oggetto digitale nello stato in cui viene correntemente conservato nel deposito.

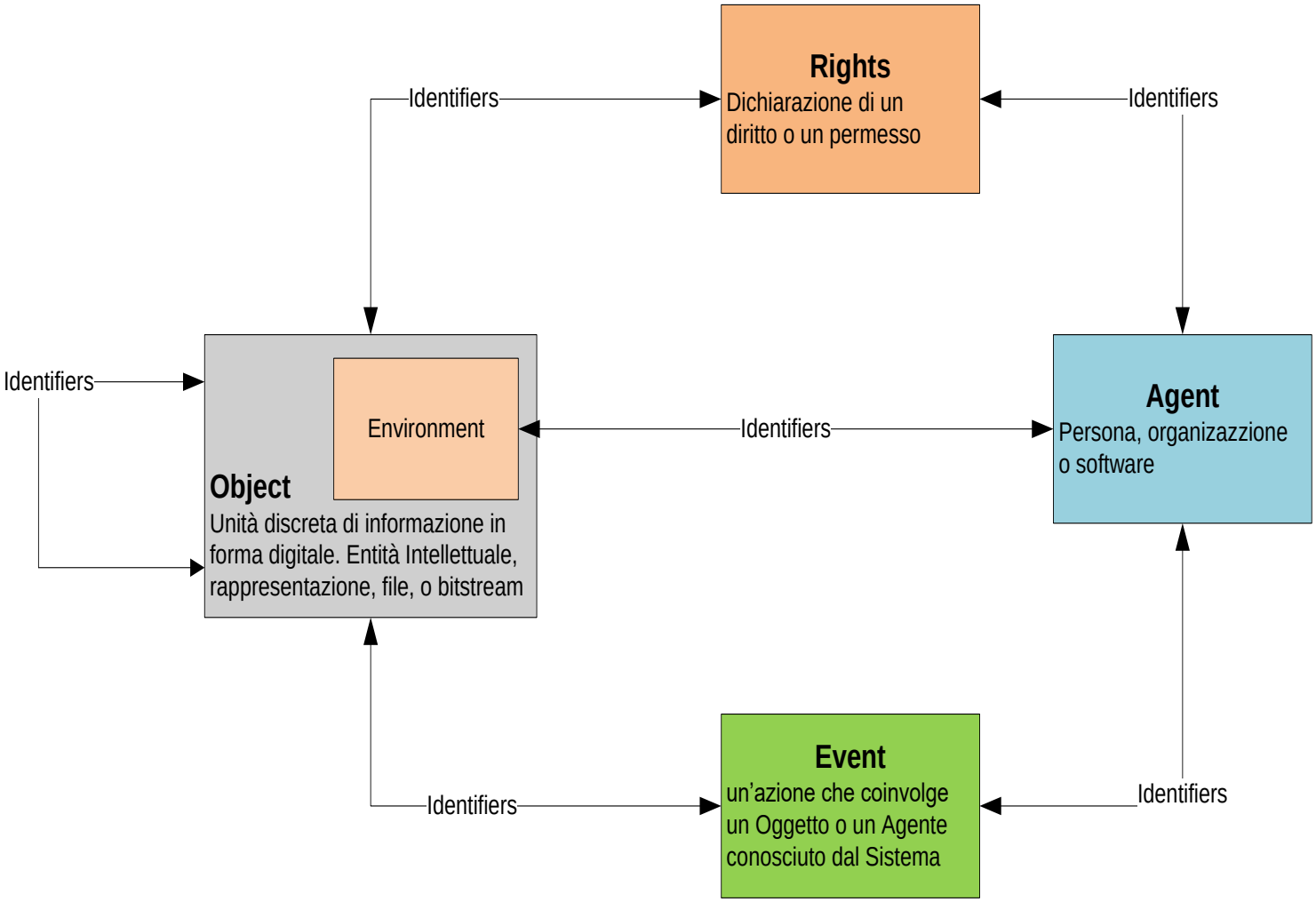
- **Gestione dei diritti**: qualsiasi diritto connesso e che possa limitare i poteri del deposito di intraprendere azioni per preservare l'oggetto digitale e per rendere accessibile l'oggetto agli utenti attuali e futuri.

PREMIS

La versione 3 dei metadati PREMIS prevede **5 entità**:

- **Oggetto (Object),** aggrega informazioni su un oggetto digitale (definito come unità discreta di informazione in forma digitale) gestito da un deposito di conservazione e ne descrive le caratteristiche rilevanti ai fini conservativi.
- **Ambiente** (**Environment**), le tecnologie che supportano la vita di un oggetto digitale: consistono di software, hardware o una loro combinazione.
- **Evento (Event),** un'azione legata alla conservazione digitale che coinvolga almeno un oggetto e/o un agente.
- **Agente (Agent),** una persona, un'organizzazione, o un software associato agli eventi di conservazione (svolti sulla base di diritti) durante la vita di un oggetto.
- **Diritti (Rights),** uno o più diritti o permessi legati ad un oggetto e/o ad un agente.

PREMIS


Rights
Dichiarazione di un diritto o un permesso
Identifiers
Identifiers
Identifiers
Environment
Object
Unità discreta di informazione in forma digitale. Entità Intellettuale, rappresentazione, file, o bitstream
Identifiers
Agent
Persona, organizzazione o software
Identifiers
Event
un'azione che coinvolge un Oggetto o un Agente conosciuto dal Sistema
Identifiers

PREMIS

L'entità Object ha **4 sotto-categorie**: *Intellectual Entity, Representation , File,* e *Bitstream*.

Una **Intellectual Entity** è una specifica creazione intellettuale o artistica considerata rilevante per la conservazione digitale da parte di una comunità designata. Ad esempio, un particolare libro, documento, mappa, fotografia, database, hardware o software.

Una Intellectual Entity può includere altre Intellectual Entities. Ad esempio, un sito web include una pagina web che include un'immagine. Una Intellectual Entity può avere una o più representations, digitali o non-digitali.

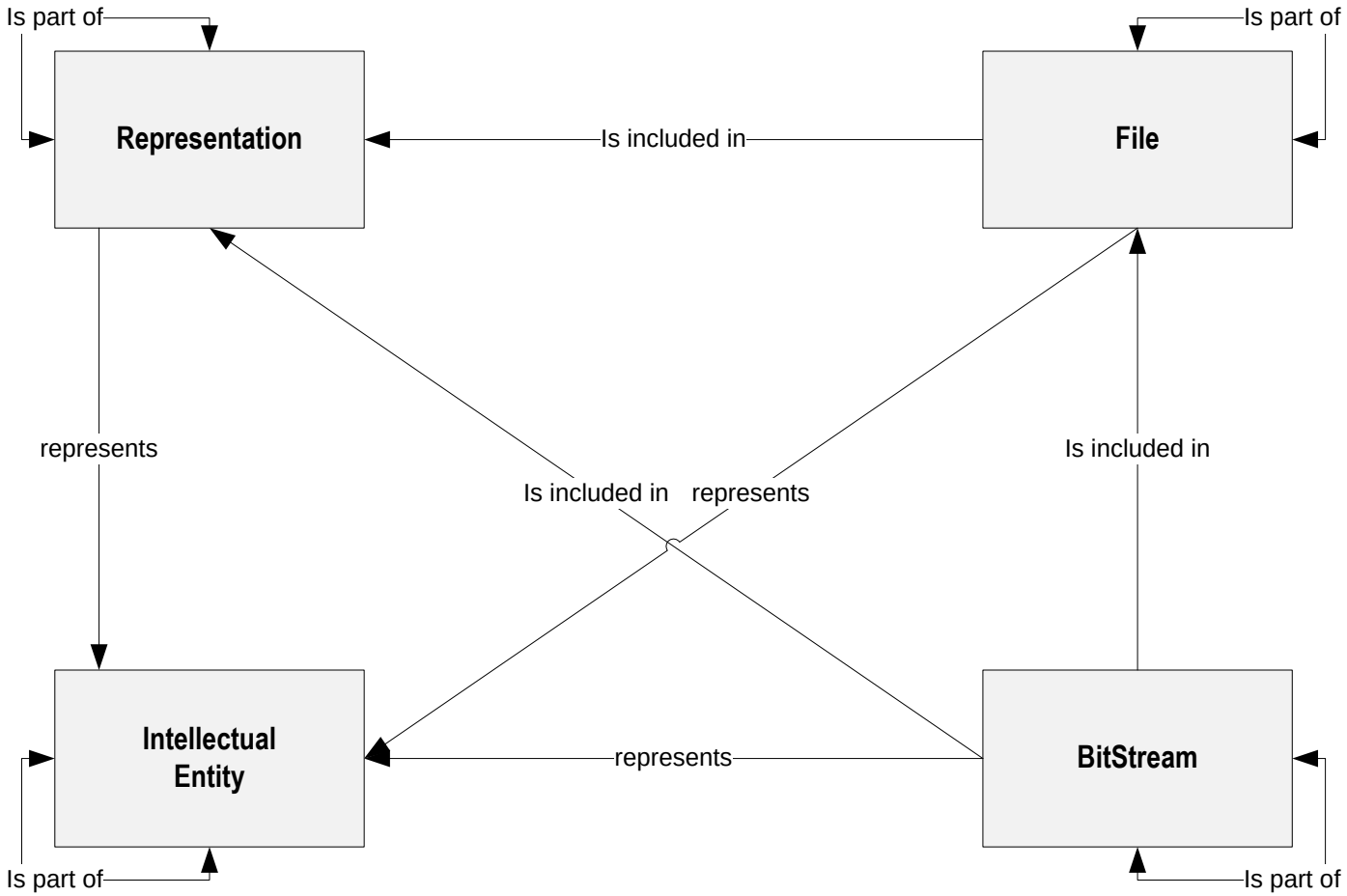
PREMIS

FILE: un sequenza ordinata di zero o più byte a cui viene attribuito un nome, riconosciuta da un sistema operativo ed accessibile dalle applicazioni. Ogni file ha un formato, definito come una struttura specifica prestabilita di un file di computer e che indica come i dati sono organizzati.

BITSTREAM: insieme di dati all'interno di un file che non può essere trasformato in un singolo file senza l'aggiunta di una struttura (intestazione, corpo ecc.) e/o riformattato per essere conforme a un particolare formato di file.

REPRESENTATION: l'insieme dei file necessari a fornire una completa e ragionevole resa di un 'entità intellettuale. Può essere pensata come la "materializzazione" digitale di un 'entità intellettuale.

PREMIS


Is part of
Representation
Is included in
File
Is part of
represents
Is included in
represents
Is included in
Intellectual Entity
represents
BitStream
Is part of
Is part of

Metadati Unisincro (UNI 11386:2010)

Lo standard UNI SinCRO, "***Supporto all'interoperabilità nella conservazione e nel recupero degli oggetti digitali***" (UNI 11386:2010) ha il compito di individuare gli elementi informativi indispensabili alla creazione dell'indice di conservazione (o "file di chiusura"), descrivendone sia la semantica sia l'articolazione sotto forma del linguaggio XML.

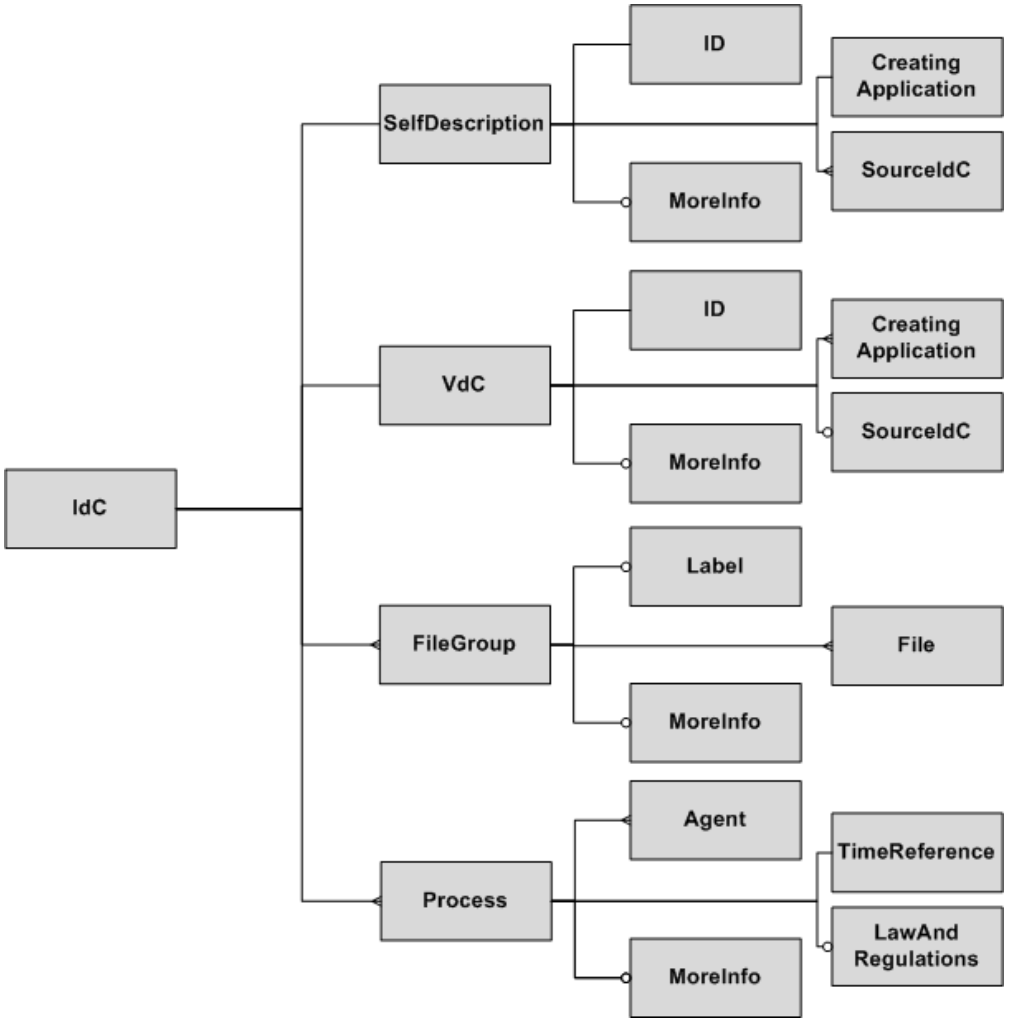
Lo standard si propone di offrire una struttura dati condivisa per favorire l'interoperabilità nei processi di migrazione rendendo possibile lo scambio di documenti tra produttori diversi.

Lo standard è indipendente da qualunque fornitore e da specifiche applicazioni.

Metadati Unisincro (UNI 11386:2010)

La struttura dell'indice di conservazione si compone di quattro elementi dipendenti direttamente dall'elemento radice <**IdC**>: **<SelfDescription>**, **<VdC>**, **<FileGroup>** e **<Process>**, .

Questi elementi aggregano degli elementi informativi specifici.



Metadati Unisincro (UNI 11386:2010)

Per **volume di conservazione** (VdC) si intende l'unità logica elementare che costituisce il risultato finale di un processo di conservazione. Si compone di uno o più file (da conservare); dell'indice di conservazione (IdC) e degli eventuali indici di conservazione pregressi.

L'**indice di conservazione** è l'evidenza informatica correlata ad ogni volume di conservazione, con un insieme di informazioni relative ai file oggetto di conservazione sostitutiva e al processo stesso, e corredata da riferimento temporale e firma digitale dei soggetti incaricati di attuare l'attività di conservazione.

Si tratta di un set di metadati definito volutamente limitato, per poter ottenere la maggior condivisione possibile, ma è implementabile estensioni per adattarlo alle esigenze di un dominio o di una comunità di operatori.

Metadati Unisincro (UNI 11386:2010)

L'elemento <**SelfDescription**> memorizza le informazioni relative all'Indice di Conservazione stesso.

L'elemento **<VdC>** (informazioni relative al Volume di Conservazione) e l'elemento **<FileGroup>** è un elemento ripetibile che riporta le aggregazioni dei file oggetto di conservazione sostitutiva.

L'elemento **<Process>** individua le informazioni sulle modalità di svolgimento del processo di conservazione sostitutiva.

Caratteristiche dei formati per la conservazione

Caratteristiche generali dei formati :

1. Apertura

Un formato è “aperto” quando è conforme a specifiche pubbliche definite da produttori, consorzi o organismi di standardizzazione riconosciuti (ISO, ETSI, Uni, etc etc).

2. Sicurezza

La sicurezza di un formato dipende da due elementi: il grado di modificabilità del contenuto del file e la capacità di essere immune dall’inserimento di codice maligno.

3. Portabilità

Si intende la facilità con cui i formati possano essere usati su piattaforme diverse, sia dal punto di vista dell’hardware che del software, inteso come sistema operativo.

Caratteristiche dei formati per la conservazione

4. Funzionalità

La possibilità da parte di un formato di essere gestito da prodotti informatici, che prevedono una varietà di funzioni messe a disposizione dell'utente per la formazione e gestione del documento informatico.

5. Supporto allo sviluppo

E' la modalità con cui si mettono a disposizione le risorse necessarie alla manutenzione e sviluppo del formato e i prodotti informatici che lo gestiscono .

6. Diffusione

La diffusione è l'impiego di uno specifico formato per la formazione e la gestione dei documenti informatici.

Documenti informatici

La possibilità di assimilare sia un database che l'estrazione di contenuti da un database ad un documento informatico è data dalle modalità di formazione regolate dall'art. 3, del DPCM 13 novembre 2014:

- *Le informazioni ed i dati conservati nel database sono prodotti sia a partire da "registrazione informatica delle informazioni risultanti da transazioni o processi informatici o dalla presentazione telematica di dati attraverso moduli o formulari resi disponibili all'utente". (art. 3, comma 1, lett. c)*
- *Il documento informatico, infine, può essere prodotto tramite la generazione o il raggruppamento anche in via automatica di un insieme di dati o registrazioni, provenienti da una o più basi dati, anche appartenenti a più soggetti interoperanti, secondo una struttura logica predeterminata e memorizzata in forma statica (art. 3, comma 1, lett. d).*

Documenti informatici

I documenti informatici possono essere prodotti anche con una estrazione statica dei dati a partire da un database.

I documenti informatici devono possedere caratteristiche di immodificabilità e di integrità determinate dalle operazioni di registrazione dell'esito delle operazioni, dall'applicazione di misure per la protezione dell'integrità delle basi di dati, dalla produzione e conservazione dei log di sistema

Nel caso di documenti amministrativi informatici, le caratteristiche di immodificabilità e di integrità possono essere ottenute anche con la loro registrazione nel registro di protocollo, in registri, in repertori, in albi, in elenchi, in archivi o raccolte di dati contenute nel sistema di gestione informatica dei documenti.

Documenti informatici

L'estrapolazione di un documento informatico da un database ricade nei casi indicati dalla norma in quanto generabile sia come elaborazione di campi e form sia come elaborazione di dati presenti in un database.

Al termine della sua formazione, il database e documento informatico e quindi il database e dovrà possedere le seguenti cinque caratteristiche fondamentali:

1. Autenticità
2. Integrità;
3. Affidabilità;
4. Leggibilità;
5. Reperibilità;

Documenti informatici

Autenticità: (caratteristica che fornisce la garanzia che il documento sia ciò che dichiara di essere, senza avere subito alterazioni o modifiche. Insieme di identificazione , provenienza e integrità);

- Identificazione univoca e garanzia dell'integrità dei singoli oggetti digitali
- Descrizione della fonte di provenienza con compilazione di metadati che descrivano il soggetto o i soggetti che hanno prodotto il database, che ne hanno contribuito alla sua compilazione nel tempo, che ne sono responsabili come custodi e gestori (continuità della conservazione)
- Generazione di metadati formato PREMIS e EAC-CPF per riportare un numero maggiore di informazioni come gli agenti e gli eventi che hanno operato sul database.

Documenti informatici

Integrità (la qualità di un documento di essere completo e inalterato, cioè non avere subito modifiche non autorizzate;) :

- Staticizzazione dei contenuti per avere la garanzia che il database conservato non abbia avuto modifiche o alterazioni rispetto alla sua forma originaria e che non siano presenti elementi che possono modificarne il contenuto dinamicamente (cancellazione di viste, codici, store procedure, etc etc);
- Documentazione delle azioni compiute sulla struttura e sulla dinamica del database originario;
- Memorizzazione di tutte le azioni sui metadati PREMIS collegati al pacchetto di versamento UNISINCRO.

Documenti informatici

Affidabilità (esprime il livello di fiducia che l'utente, cioè colui che legge il documento ripone, o può riporre nel documento informatico, in particolare nella sua visualizzazione leggibile allo stesso)

- Documentazione della fonte e delle caratteristiche di produzione del database
- Documentazione delle azioni compiute (o non compiute) sui contenuti del database originale;
- Documentazione delle caratteristiche di storicizzazione dei contenuti del database
- Memorizzazione di tutte le azioni sui metadati PREMIS collegati al pacchetto di versamento descritto in UNISINCRO.

Documenti informatici

Leggibilità (a caratteristica che definisce il mantenimento della fruibilità delle informazioni contenute nel documento durante l'intero ciclo di gestione dei documenti dalla formazione alla conservazione)

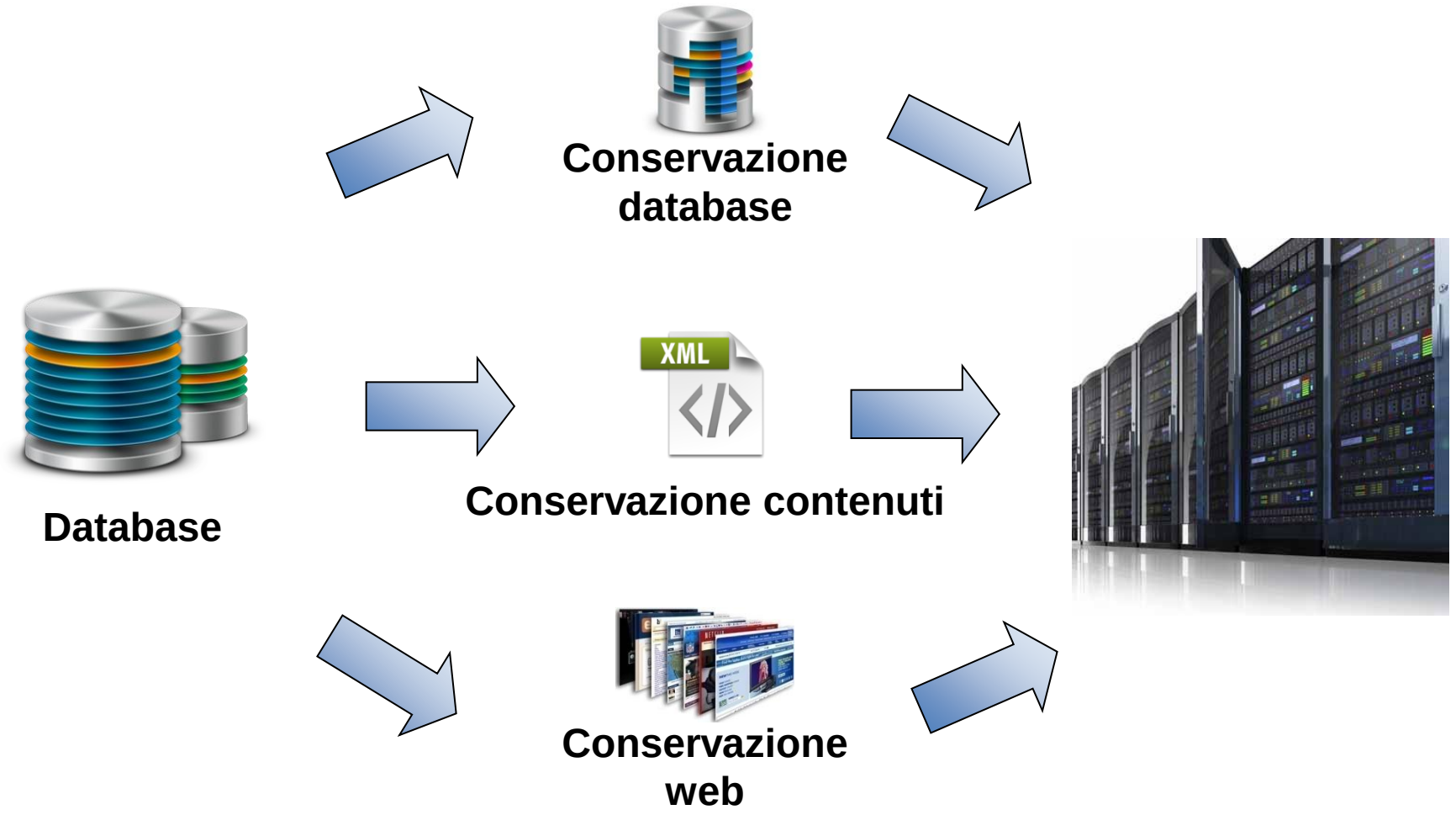
- Creazione di una check list della documentazione disponibile e Inserimento della documentazione descrittiva del database e del suo uso nel tempo;
- Utilizzo del formato SIARD per conservare e rendere leggibile un database ed i suoi contenuti e del formato WARC per conservare i contenuti web
- Utilizzo completo dei metadati del formato SIARD per i dati del processo e dei contenuti del database archiviato;
- Memorizzazione di tutte le azioni sui metadati PREMIS collegati al pacchetto di versamento descritto in UNISINCRO;

Documenti informatici

Reperibilità (esprime la capacità di reperire ed esibire il documento con le caratteristiche di leggibilità, integrità, affidabilità, autenticità).

- Utilizzo del formato SIARD per conservare e rendere leggibile un database ed i suoi contenuti e del formato WARC per conservare i contenuti web
- Memorizzazione di tutte le azioni sui metadati PREMIS collegati al pacchetto di versamento descritto in UNISINCRO;

Scenari


Conservazione database
Database
XML
Conservazione contenuti
Conservazione web

Scenario 1 : sintesi processo di conservazione

1. **Preparazione del processo e documentazione a corredo del database**

   (check list, conversione PDF/A2, identificazione e generazione HASH)

2. **Accesso**

   2.1 Fornitura o accesso al database

   2.2 Normalizzazione e conversione database

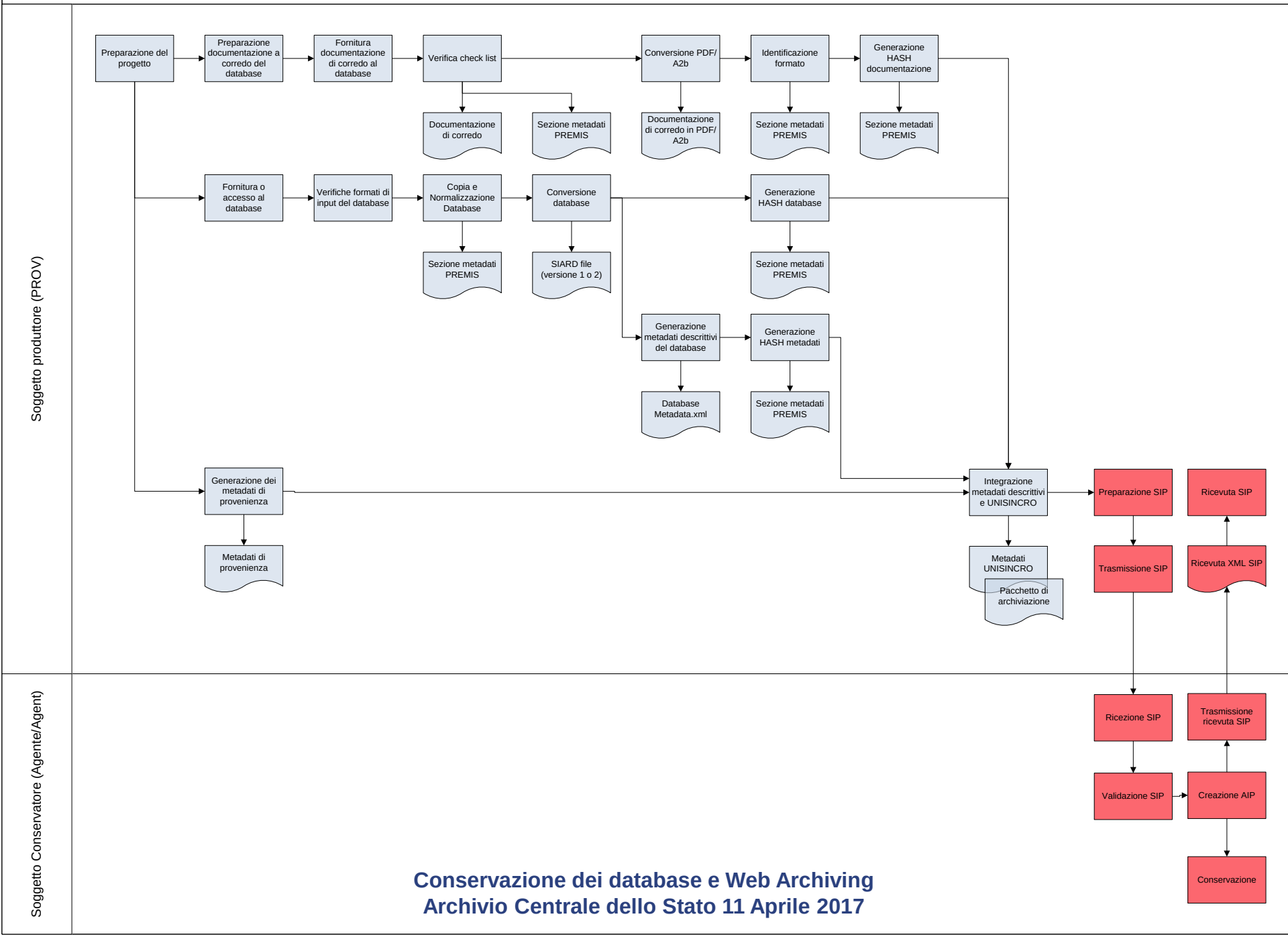
   2.3 Generazione metadati descrittivi del database

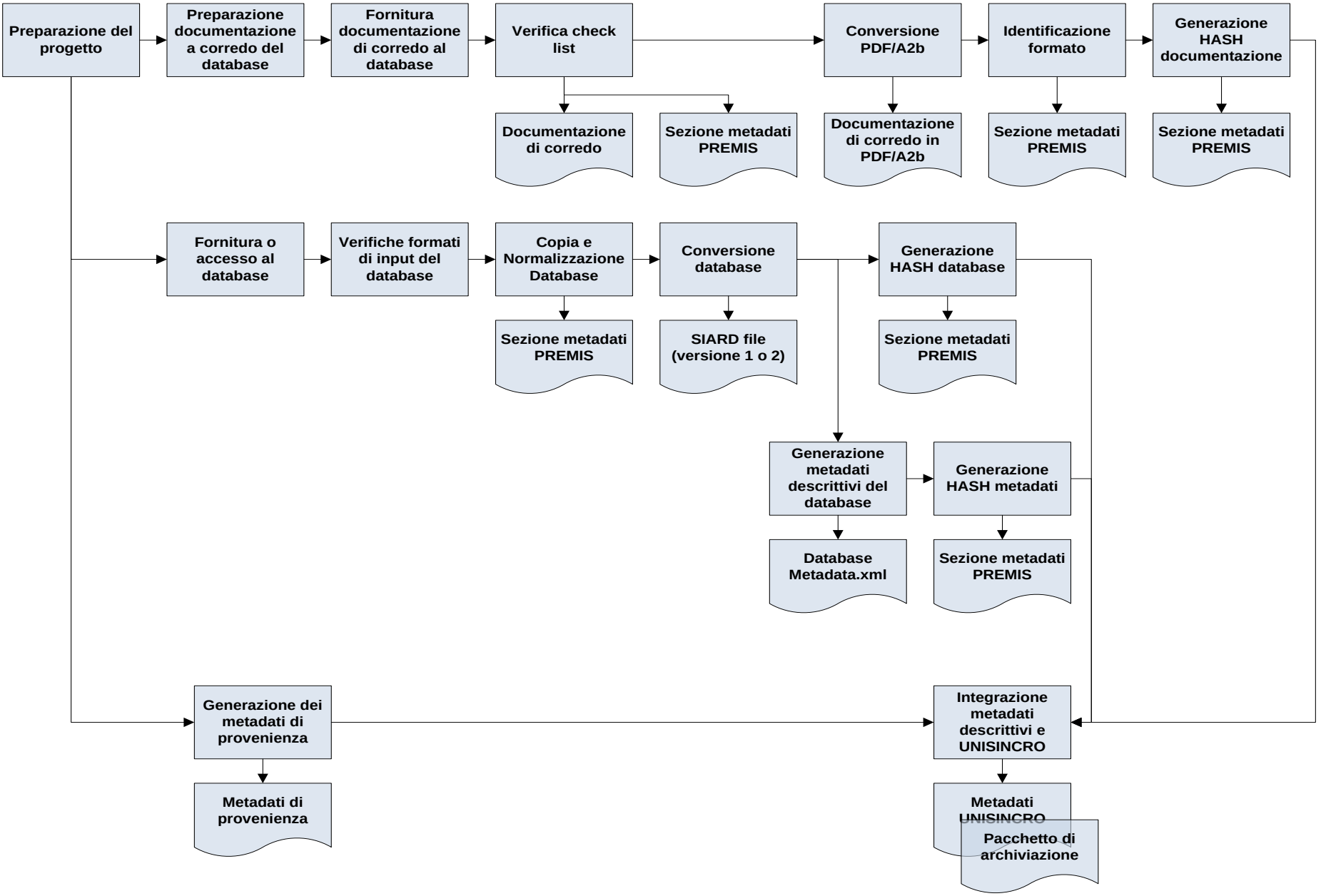
   2.4 Generazione HASH database e metadati descrittivi

3. **Metadati**

   3.1 Generazione dei metadati di provenienza

   3.2 Integrazione dei metadati descrittivi e UNISINCRO


Soggetto produttore (PROV)
Preparazione del progetto
Preparazione documentazione a corredo del database
Fornitura documentazione di corredo al database
Verifica check list
Documentazione di corredo
Sezione metadati PREMIS
Conversione PDF/ A2b
Documentazione di corredo in PDF/ A2b
Identificazione formato
Sezione metadati PREMIS
Generazione HASH documentazione
Sezione metadati PREMIS
Fornitura o accesso al database
Verifiche formati di input del database
Copia e Normalizzazione Database
Sezione metadati PREMIS
Conversione database
SIARD file (versione 1 o 2)
Generazione HASH database
Sezione metadati PREMIS
Generazione metadati descrittivi del database
Database Metadata.xml
Generazione HASH metadati
Sezione metadati PREMIS
Generazione dei metadati di provenienza
Metadati di provenienza
Integrazione metadati descrittivi e UNISINCRO
Metadati UNISINCRO
Pacchetto di archiviazione
Preparazione SIP
Trasmissione SIP
Ricevuta SIP
Ricevuta XML SIP
Soggetto Conservatore (Agente/Agent)
Ricezione SIP
Validazione SIP
Creazione AIP
Trasmissione ricevuta SIP
Conservazione
Conservazione dei database e Web Archiving
Archivio Centrale dello Stato 11 Aprile 2017


Preparazione del progetto
Preparazione documentazione a corredo del database
Fornitura documentazione di corredo al database
Verifica check list
Conversione PDF/A2b
Identificazione formato
Generazione HASH documentazione
Documentazione di corredo
Sezione metadati PREMIS
Documentazione di corredo in PDF/A2b
Sezione metadati PREMIS
Sezione metadati PREMIS
Fornitura o accesso al database
Verifiche formati di input del database
Copia e Normalizzazione Database
Conversione database
Generazione HASH database
Sezione metadati PREMIS
SIARD file (versione 1 o 2)
Sezione metadati PREMIS
Generazione metadati descrittivi del database
Generazione HASH metadati
Database Metadata.xml
Sezione metadati PREMIS
Generazione dei metadati di provenienza
Integrazione metadati descrittivi e UNISINCRO
Metadati di provenienza
Metadati UNISINCRO
Pacchetto di archiviazione

Scenario 1 : processo di conservazione

Preparazione e fornitura della documentazione a corredo del database

Predisposizione della documentazione disponibili che può essere utile nel futuro per comprendere la struttura del database originario e l'utilizzo che ne è stato fatto.

Verifica check list

Documentazione utente dell'applicazione
Documentazione tecnica dell'applicazione
Documentazione tecnica del database
Schema logico database
Schema fisico database
Descrizione del database nel Data Definition Language

Elenco singolo delle tabelle
Elenco singolo delle viste
Elenco singolo dei trigger
Elenco singolo delle procedure
Documento riassuntivo per l'intero database con tabelle, viste, trigger e procedure.
Descrizione del database nel Data Definition Language

Check list documentazione e Premis

La documentazione prevista dalla check list è riportata nei metadati PREMIS inserendo per ogni oggetto la coppia:

premis:significantPropertiesType= Elenco singolo delle viste, Elenco singolo dei trigger, Elenco singolo delle procedure, etc etc

premis:significantPropertiesValue= ID Object PREMIS (identificativo PREMIS dell'oggetto

---

```
<premis:significantProperties>
<premis:significantPropertiesType>Elenco singolo delle viste</premis:significantPropertiesType>
<premis:significantPropertiesValue>WA_ID_00000026</premis:significantPropertiesValue>
</premis:significantProperties>
<premis:significantProperties>
<premis:significantPropertiesType>Schema fisico database</premis:significantPropertiesType>
<premis:significantPropertiesValue>WA_ID_00000022</premis:significantPropertiesValue>
</premis:significantProperties>
```

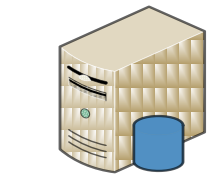
Eventi di normalizzazione del database

Tutte le azioni di normalizzazione che non intervengono sui contenuti essenziali vengono conservate come eventi specifici relativi ad un oggetto nei metadati di conservazione PREMIS.

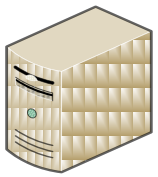
Gli eventi eseguiti sul database e monitorato sono stati definiti come:

- Database Normalization: remove user (cancellazione degli utenti del database per mantenere solo l'amministratore);
- Database Normalization: remove table (cancellazione delle tabelle vuote o non funzionali alla conservazione);
- Database Normalization: remove view (rimuove anche stored procedure e funzioni SQL Server).

Estrazione dei contenuti e generazione file SIARD



Database Server



Operation Server

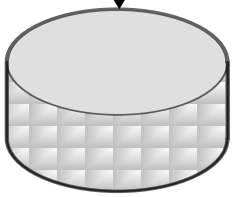
Collegamento con il database di origine

SQL Server

Porta 1433

Database Preservation Toolkit

Metadati XML del Database SIARD



Database SIAS ARCHIVISTA

Database in formato SIARD

SIARD

SIARD (Software Independent Archiving of Relational Databases) è un formato aperto appositamente sviluppato per l'archiviazione di banche dati relazionali

Il formato possiede le caratteristiche più rilevanti per la conservazione dei contenuti di un database nel tempo:

- conserva le informazioni, non la loro rappresentazione o interazioni;
- conserva i dati primari, non il codice;
- conserva le tabelle e le loro relazioni;

I vincoli non sono conservati in quanto si presume che i database da conservare sono già consistenti e che non verranno più modificati. Altri elementi come Triggers, Stored Procedures non sono archiviati in quanto non è possibile garantire la leggibilità di questi elementi indipendentemente dal motore di database utilizzato.

SIARD

La memorizzazione in formato SIARD garantisce una durata di vita molto più lunga dei dati rispetto a qualsiasi altro formato proprietario non standardizzato.

La struttura del file SIARD segue l'approccio di formati moderni di tipo contenitore che contiene al suo interno file xml, file di testo e file binari (come MS Office Open XML o come Open Document Format).

Il file ZIP è a 64-bit per poter gestire file di dimensione maggiore di 4 GB ed è utilizzato come contenitore senza compressione e tutti i file XML sono memorizzati con codifica UTF-8.

Ogni tabella è numerata progressivamente a partire da table0 ed ogni campo di ogni singola tabella ha nome c1,c2, - - - -, cx.

SIARD

La memorizzazione in formato SIARD garantisce una durata di vita molto più lunga dei dati rispetto a qualsiasi altro formato proprietario non standardizzato.

La struttura del file SIARD segue l'approccio di formati moderni di tipo contenitore che contiene al suo interno file xml, file di testo e file binari (come MS Office Open XML o come Open Document Format).

Il file ZIP è a 64-bit per poter gestire file di dimensione maggiore di 4 GB ed è utilizzato come contenitore senza compressione e tutti i file XML sono memorizzati con codifica UTF-8.

Ogni tabella è numerata progressivamente a partire da table0 a tablex ed ogni campo di ogni singola tabella ha come nome c1,c2, - - - -, cx.

Le informazioni per la decodifica originale del nome sono nei metadati descrittivi.

SIARD: sostenibilità come formato

Il formato SIARD può essere considerato adatto e compatibile per un processo di conservazione in quanto:

- è "**non proprietario**" e non è sottoposto a restrizioni come licenze o/e brevetti;
- è "**aperto**": le specifiche del formato sono liberamente disponibili e documentate;
- è "**standard**": è basato su di una serie di standard nazionali ed internazionali (eCH-0165, ISO/IEC 9075:2008, ISO/IEC 10646:2012, XML);
- è "**trasparente**": è un formato contenitore per i contenuti di un database;
- non ha meccanismi tecnici di protezione dei contenuti;
- auto-documentato in quanto ciascuna componente interna (tabelle e campi) ha propri metadati descrittivi.

SIARD 2

L'ambito del progetto E-ARK è stata realizzata la versione 2.0 di SIARD che include una serie di miglioramenti e potenzialità aggiuntive rispetto alla prima versione:

- il supporto per lo standard SQL:2008 (ISO/IEC 9075), comprensiva dei tipi di dati, array e tipi dati definiti dagli utenti;
- l'utilizzo di regole di validazione sia per la struttura che per i contenuti nei file XML;
- salvataggio dei Large OBject (LOB) inclusi nel database insieme agli altri contenuti
- Supporto per il metodo di compressione “deflate”
- Retro compatibilità con la versione 1.0.

https://it.wikipedia.org/wiki/Deflate

ISTITUTO CENTRALE PER GLI ARCHIVI

SIARD: metadati descrittivi

Il formato SIARD prevede una serie di metadati descrittivi del contenuto del file, del processo di archiviazione e di provenienza:

Nome del metadato	Descrizione	Obb
dbname	Nome del database archiviato	SI
description	Breve descrizione del contenuto del database	
archiver	Nome del responsabile dell'archiviazione del database	
archiverContact	Dati di contatto (telefono, mail) del responsabile dell'archiviazione del database	
dataOwner	Nome del proprietario dei dati del database quando questo è stato archiviato.	SI
dataOriginTimespan	Periodo temporale dei dati contenuti nel database.	SI
producerApplication	Nome e versione del programma che ha generato i metadati	
archivalDate	data di creazione dell'archivio SIARD con il database	SI
messageDigest	Message digest di tutti i dati contenuti nel folder content	SI
clientMachine	Nome della macchina nella quale è stato eseguito il programma SIARD per l'archiviazione.	
databaseProduct	name of database product and version from which database originates	
connection	Stringa di connessione utilizzata per il processo di archiviazione	
databaseUser	Nome utente del database utilizzato per il processo di archiviazione	
schemas	Lista degli schemi presenti nel database	SI
users	Lista degli utenti presenti nel database archiviato	SI
roles	Lista degli ruoli presenti nel database archiviato	
privileges	Lista degli privilegi presenti nel database archiviato	

ISTITUTO CENTRALE PER GLI ARCHIVI

Metadati di provenienza: SIARD e EAG

Il file SIARD del database ha una serie di metadati che sono estrapolati automaticamente, memorizzati in un file xml separato e resi statici con un hash specifico. In questo modo è possibile consultarli direttamente senza doverli estrarre nuovamente.

Sono stati definiti anche metadati EAG per la descrizione del conservatore che saranno inseriti nei metadati UNISINCRO.

```
<?xml version="1.0" encoding="UTF-8"?>
<eag:eag xsi:schemaLocation="http://www.archivesportaleurope.net/Portal/profiles/eag_2012/
http://www.archivesportaleurope.net/Portal/profiles/eag_2012.xsd"
xmlns:eag="http://www.archivesportaleurope.net/Portal/profiles/eag_2012/">
<eag:control xmlns="">
<eag:recordId>IT-ICAR</eag:recordId>
<eag:maintenanceAgency>
<eag:agencyCode>ICAR</eag:agencyCode>
<eag:agencyName>Istituto centrale per gli archivi</eag:agencyName>
</eag:maintenanceAgency>
<eag:maintenanceStatus>new</eag:maintenanceStatus>
<eag:maintenanceHistory>-----</eag:maintenanceHistory>
<eag:sources>
<eag:source href="http://www.icar.beniculturali.it">
<eag:sourceEntry>Istituto centrale per gli archivi Web Site</eag:sourceEntry>
</eag:source>
</eag:sources>
</eag:control>
<eag:archguide xmlns="">
<eag:identity>
<eag:autform>Istituto centrale per gli archivi</eag:autform>
<eag:autform>ICAR</eag:autform>
</eag:identity>
<eag:desc>
<eag:repositories>
<eag:repository>
<eag:geogarea>Europe</eag:geogarea>
<eag:location localType="postal address">
<eag:country>Italy</eag:country>
<eag:municipalityPostalcode>00185 Roma</eag:municipalityPostalcode>
<eag:street>Viale Castro Pretorio 105 </eag:street>
</eag:location>
<eag:webpage href="http://www.icar.beniculturali.it">Home page</eag:webpage>
------------------
<eag:descriptiveNote>--------</eag:descriptiveNote>
</eag:repository>
</eag:repositories>
</eag:desc>
</eag:archguide>
</eag:eag>
```

Metadati Premis

Gli oggetti interessati alla conservazione avranno un loro identificativo che sarà utilizzato nei metadati PREMIS ed UniSincro per la gestione delle relazioni fra le varie entità.

Oggetti

Entità	Codice identificativo	Contenuti
Object	WA_ID_00000020	Archivista_cons.siard
Object	WA_ID_00000021	Archivista_cons_metadata.xml
Object	WA_ID_00000027	Modello_Fisico_Database_Archivista_Convertito_A2b.pdf
Object	WA_ID_00000025	Modello_Fisico_Database_Archivista_Originale_A2b.pdf
Object	WA_ID_00000028	Modello_Logico_Database_Archivista_Originale_A2b.pdf
Object	WA_ID_00000026	SIAS_LINEE_GUIDA_Patrimonio_4002_A2b.pdf
Object	WA_ID_00000024	SIAS_LINEE_GUIDA_Patrimonio_V_III_Pergamene_4002_A2b.pdf
Object	WA_ID_00000023	SIAS_LINEE_GUIDA_Patrimonio_V_III_Sigilli_4002_A2b.pdf
Object	WA_ID_00000029	SIAS_LINEE_GUIDA_Patrimonio_V_I_Inventario_4002_A2b.pdf
Object	WA_ID_00000022	SIAS_Modello_dati_Amanuense_Archivista_A2b.pdf
Object	WA-S-00000001	Windows 10
Object	WA-S-00000011	DB Visualization Toolkit
Object	WA-S-00000005	Notepad 10
Object	WA-S-00000004	Firefox 47

Metadati Premis

Eventi, Agenti e Diritti

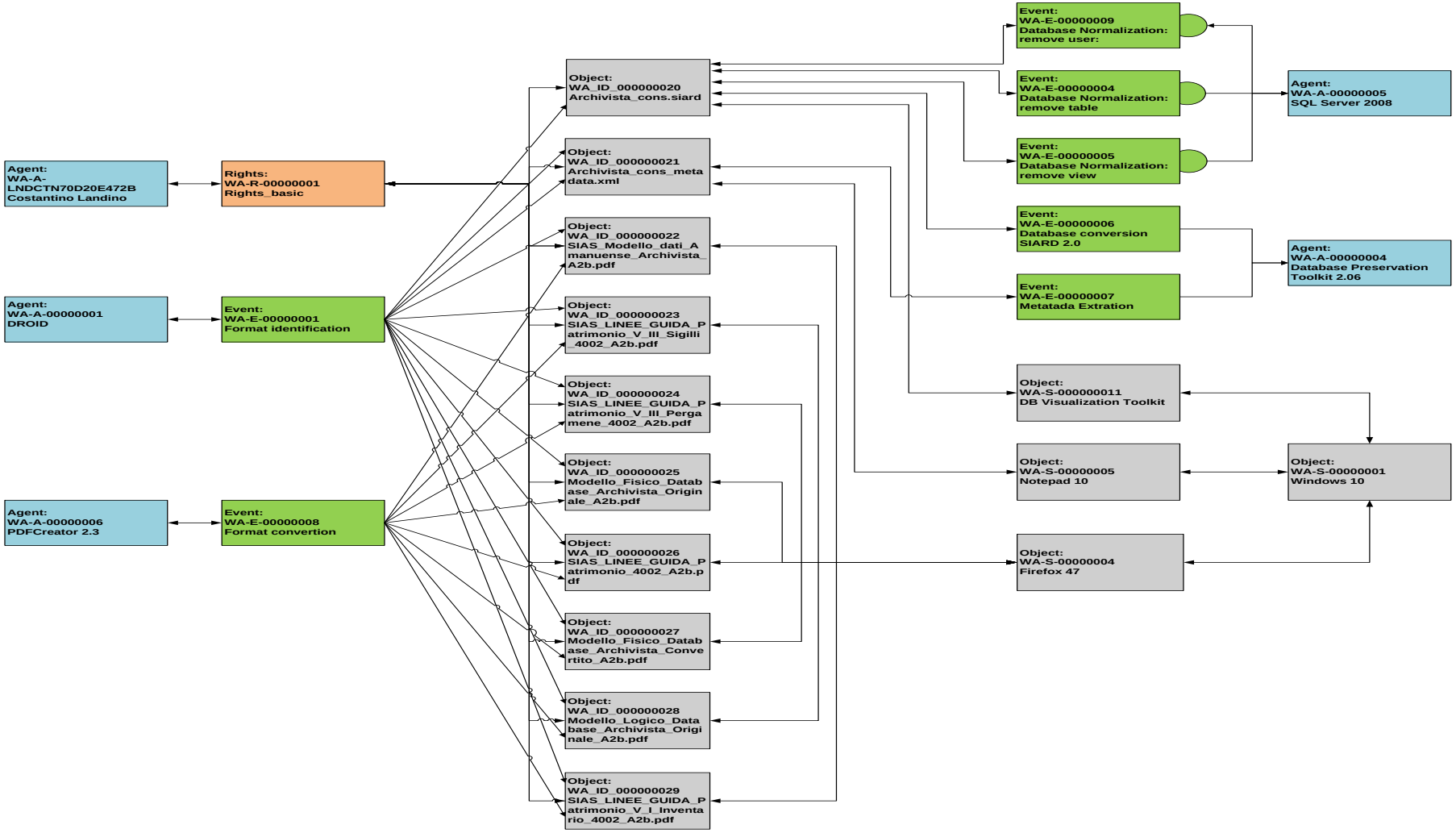
Entità	Codice identificativo	Contenuti
Events	WA-E-00000001	Format identification
Events	WA-E-00000008	Format conversion
Events	WA-E-00000006	Database conversion SIARD 2.0
Events	WA-E-00000007	Metatada extration
Events	WA-E-00000009	Database Normalization: remove user
Events	WA-E-00000004	Database Normalization: remove table
Events	WA-E-00000005	Database Normalization: remove view
Agents	WA-A-00000001	DROID
Agents	WA-A-00000006	PDFCreator 2.3
Agents	WA-A-00000005	SQL Server 2008
Agents	WA-A-00000004	Database Preservation Toolkit 2.06
Agents	WA-A-LNDCTN70D20E472B	Costantino Landino
Rights	WA-R-Rights_basic	Diritti generali di accesso

Sono stati individuati 5 agenti, 1 schema di diritti, 7 tipologie di eventi, 4 intellectual entities e 10 oggetti.

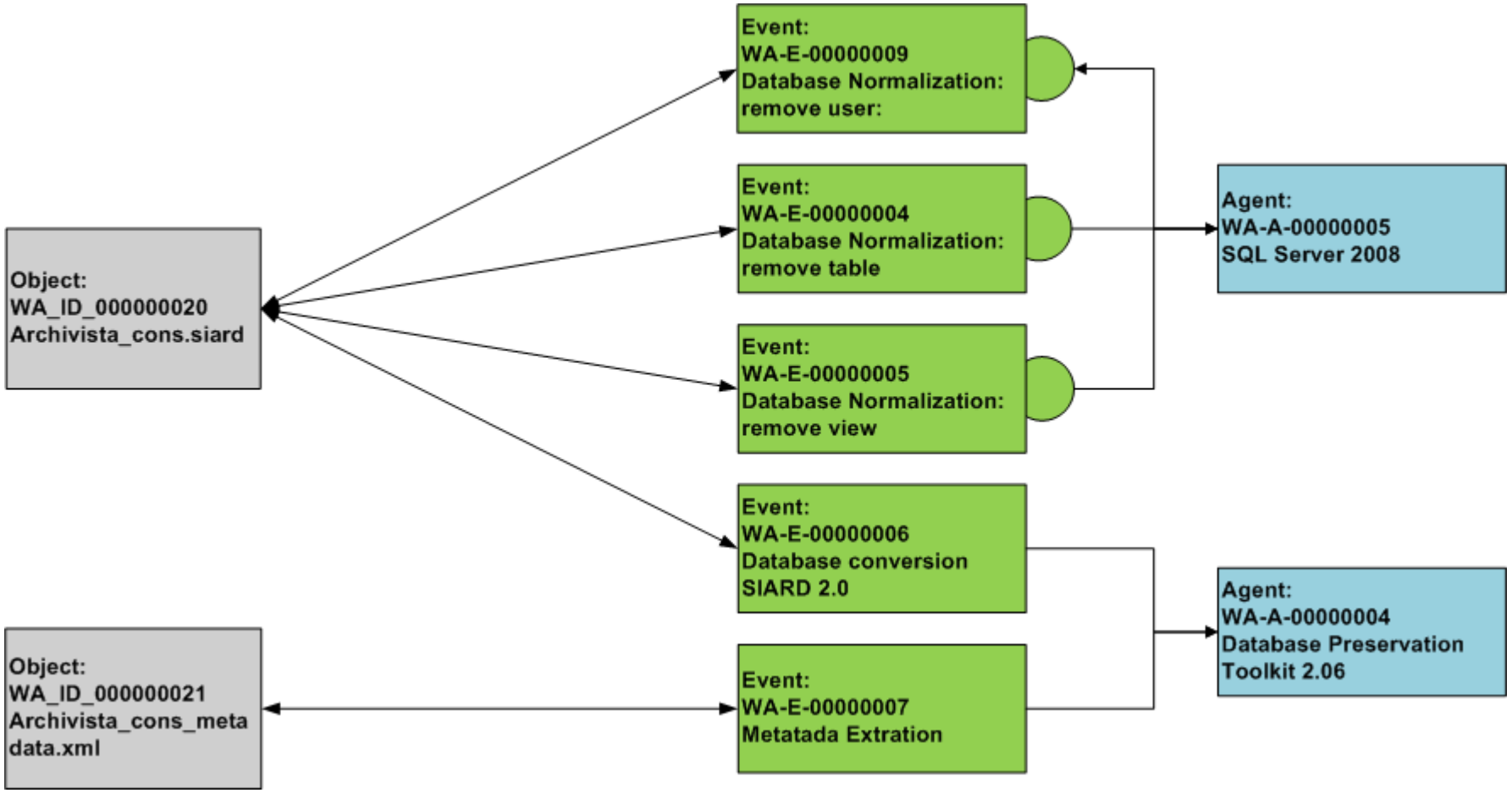
Questi creano fra di loro uno schema di relazioni molto fitto.

ISTITUTO CENTRALE PER GLI ARCHIVI

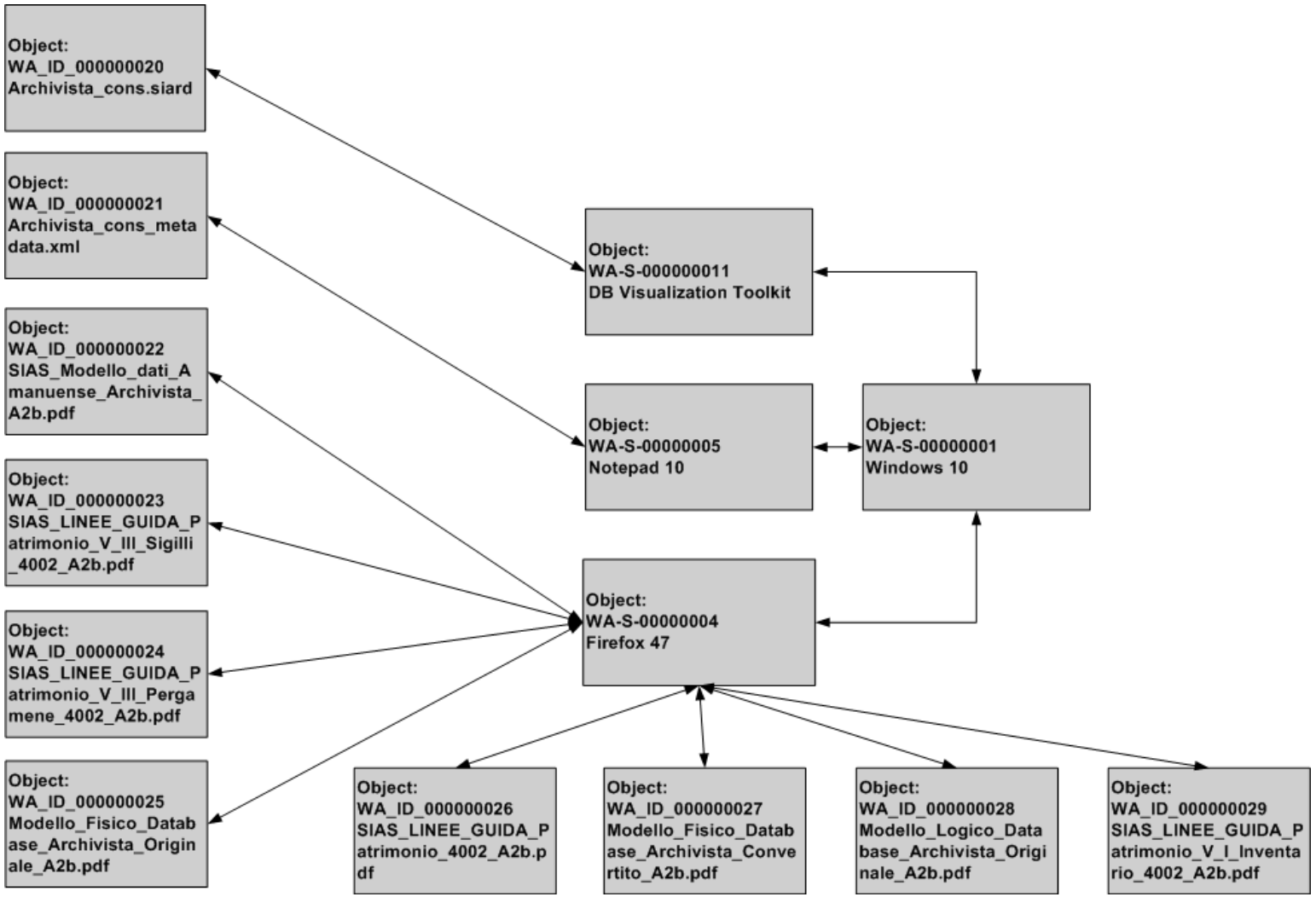
Relazioni complessive nei metadati Premis



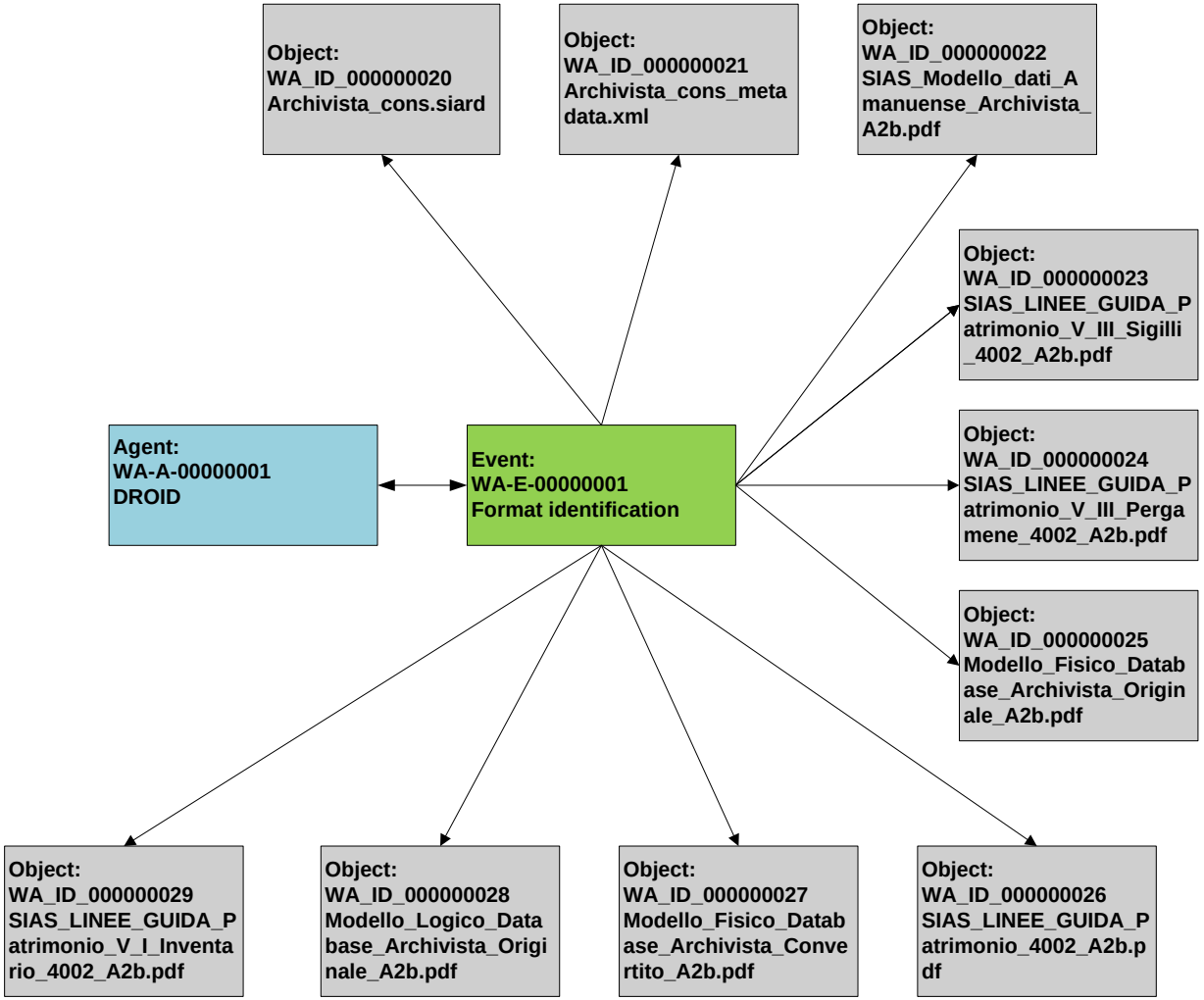
Relazioni nei metadati Premis relative al database



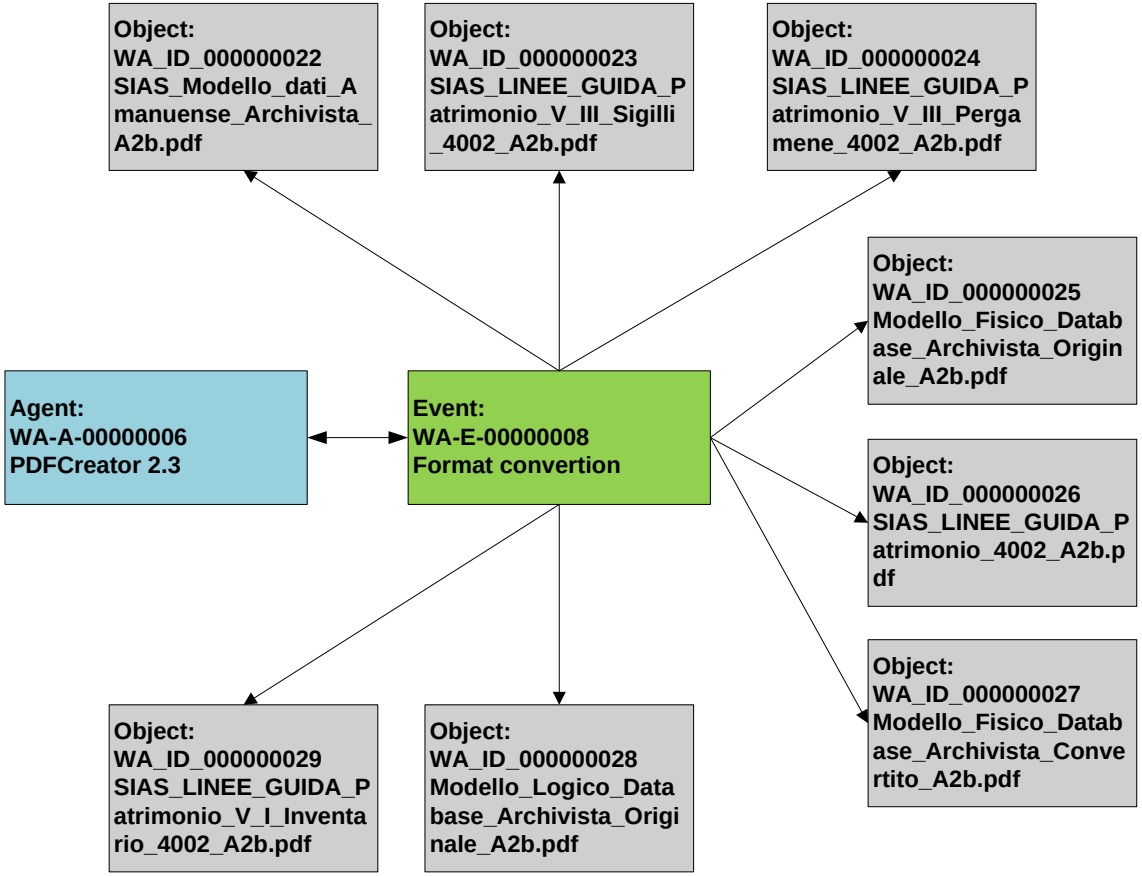
Relazioni nei metadati Premis: objects



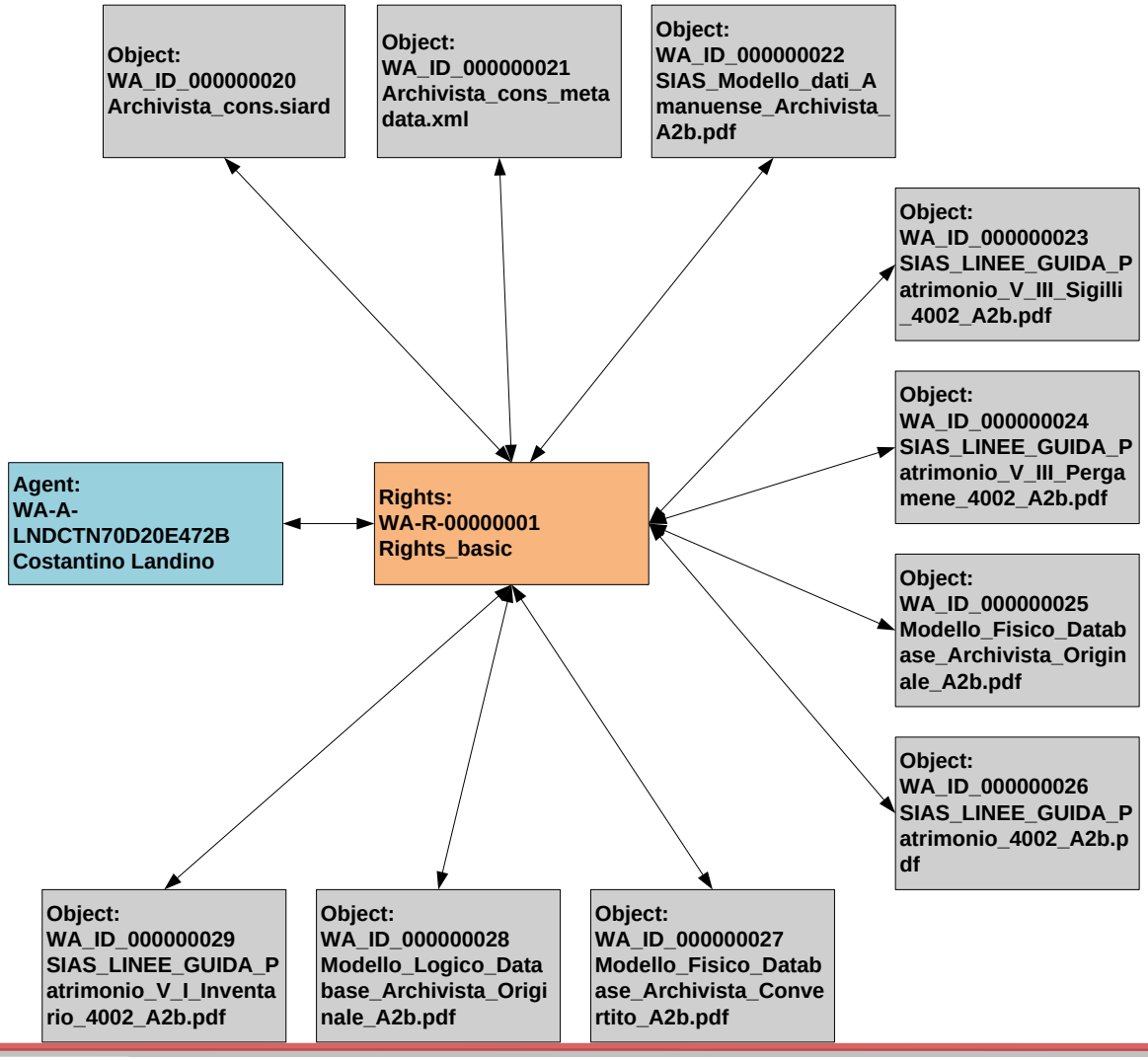
Relazioni nei metadati Premis: events



Relazioni nei metadati Premis: events



Relazioni nei metadati Premis: rights



ISTITUTO CENTRALE PER GLI ARCHIVI

Indice di conservazione

L'indice di conservazione è stato costruito a partire dai metadati PREMIS , con gli oggetti di conservazione del database SIAS, il database in formato SIARD, i metadati SIARD in formato XML e la documentazione di descrizione del database e dell'applicazione originale.

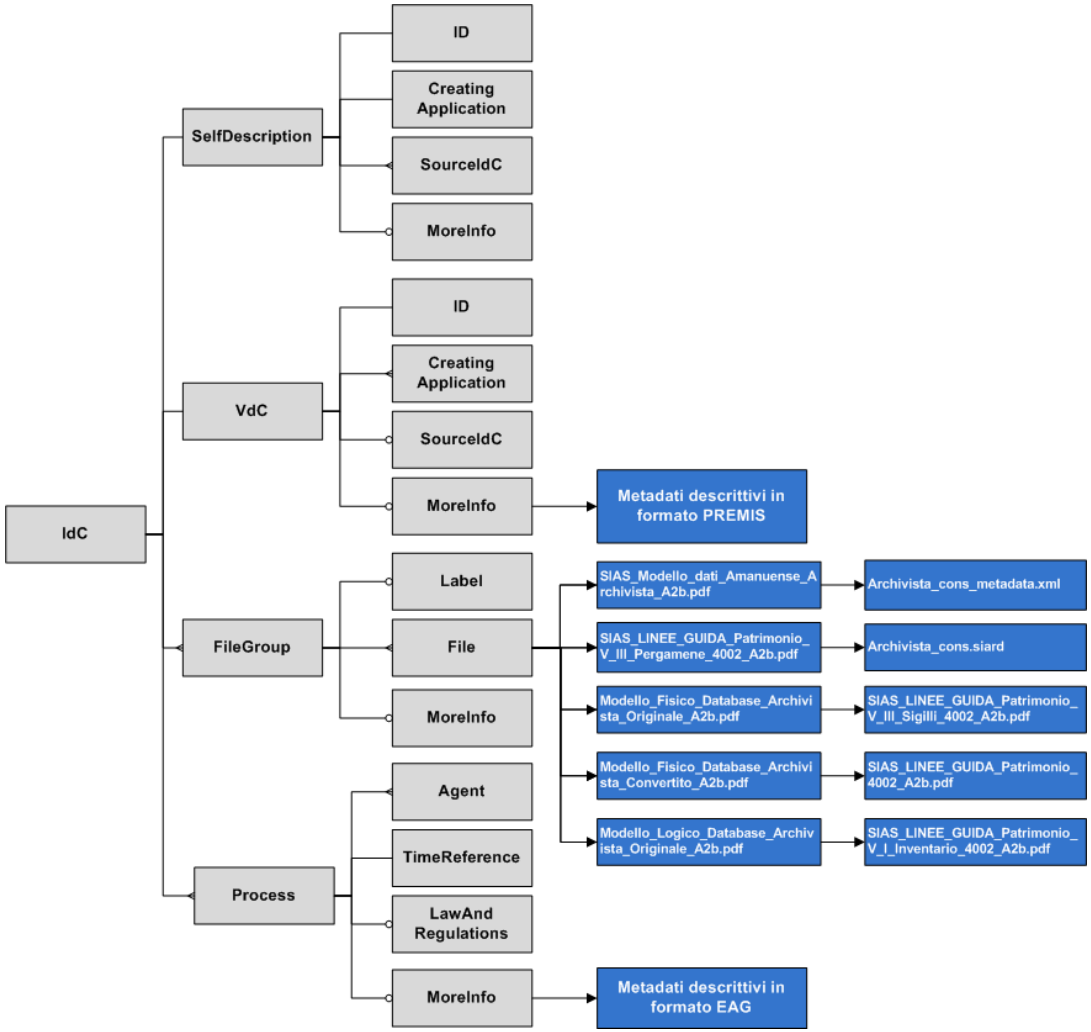
I metadati PREMIS relativi agli oggetti da sottomettere in conservazione sono stati inseriti come embedded metadata nella sezioneVdC

I metadati PREMIS relativi agli oggetti da sottomettere in conservazione sono stati inseriti come embedded metadata nella sezioneVdC

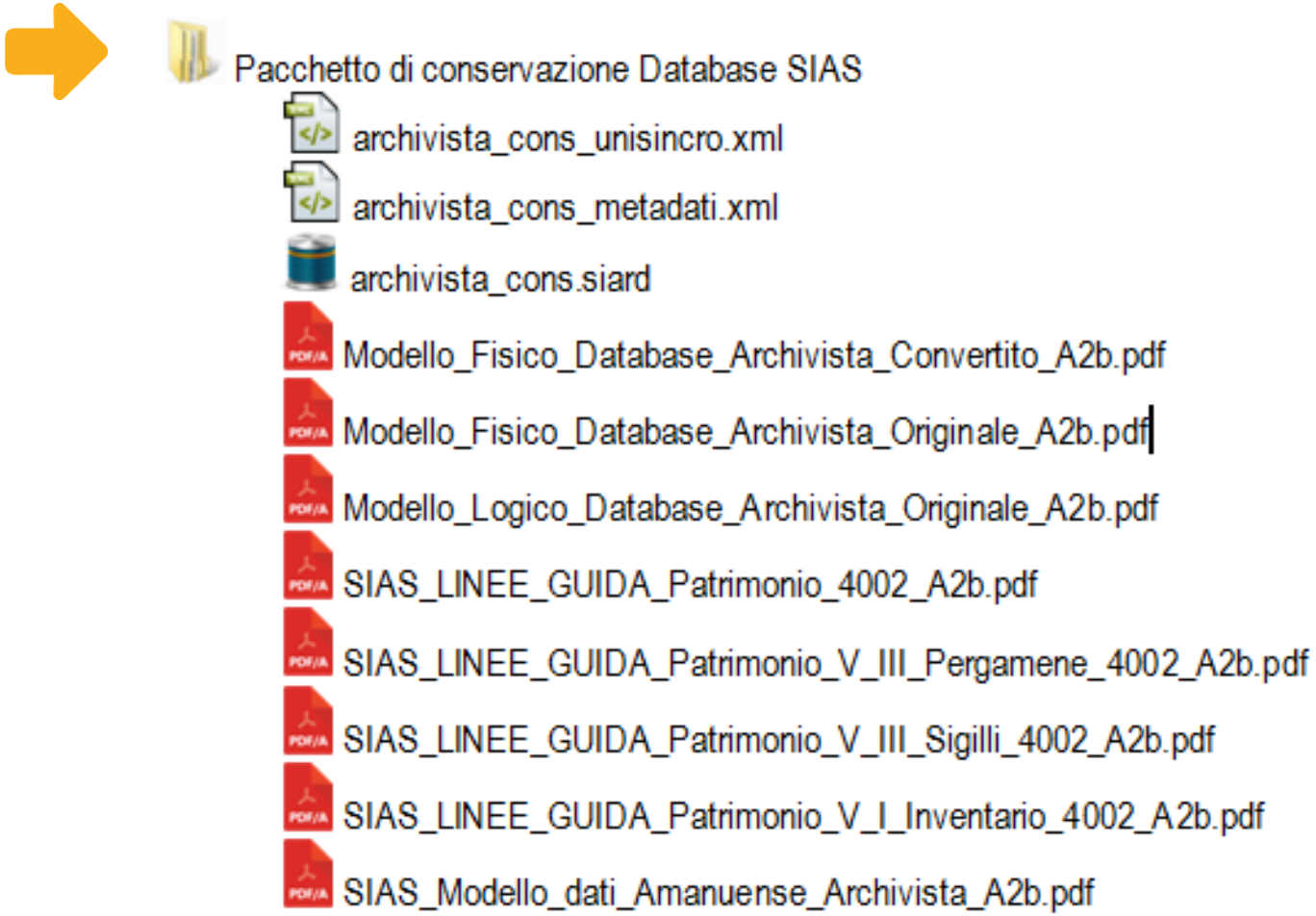
Filegroup e file contengono le informazioni sui file contenuti nel pacchetto.

I metadati descrittivi dell'ICAR in formato EAG sono stati inseriti nella sezione moreinfo di process.

Indice di conservazione



Pacchetto di conservazione Database SIAS



Strumenti: SIARD Suite

La SIARD Suite è un pacchetto di programmi per il supporto all'adozione dello standard SIARD e comprende tre applicazioni principali:

SiardFromDb è uno strumento che permette la conversione di banche dati Oracle, Microsoft SQL Server e Microsoft Access in un file nel formato archiviabile SIARD.

SiardToDb permette di caricare i file SIARD in banche dati Oracle, Microsoft SQL Server e Microsoft Access

SiardEdit permette all'utente di completare e aggiornare i metadati, di eseguire ricerche al loro interno e di visionare i dati primari.

Le applicazioni di SIARD Suite sono indipendenti da piattaforme e da prodotti software proprietari e sono disponibili gratuitamente.

Strumenti: Database Preservation Toolkit

Il Database Preservation Toolkit permette la conversione di basi di dati in tempo reale in vari formati di conservazione, come SIARD

Il toolkit permette anche la conversione da un file SIARD verso un nuovo database.

conservazione formati di nuovo in sistemi live per consentire la piena funzionalità del database.

Questo toolkit nasce nell'ambito del progetto RODA ed è stato rilasciato come un progetto autonomo per l'interesse su questa particolare problematica.

E' stato ulteriormente sviluppato nell'ambito del progetto EARK insieme ad una nuova versione del formato conservazione SIARD

Strumenti: Database Preservation Toolkit

Il Database Preservation Toolkit supporta la conversione nei formati :SIARD 1, SIARD 2 e SIARD DK dei Database Management Systems:

- MySQL/MariaDB
- PostgreSQL
- Oracle
- Microsoft SQL Server
- Microsoft Access
- JDBC

Il toolkit permette anche di riversare il contenuto dei database conservati su di dei DBMS indicati in precedenza..

Strumenti: Database Visualization Toolkit

Il Database Visualization Toolkit è un visualizzatore web per database relazionali dedicato a quelli conservati in formato SIARD 1 o SIARD2.

Utilizza SOLR come backend, e permette la ricerca, la navigazione dei contenuti e la loro l'esportazione. La gestione degli indici avviene attraverso l'utilizzo del Database Preservation Toolkit e ne integra le funzioni.

Un esempio di invocazione della procedura di caricamento:

```
java -jar "-Dfile.encoding=UTF-8"
"-Ddbvtk.workspace=C:\software\dbvtk\dbvtk-data"
"C:\software\dbvtk\dbptk-app.jar"
-e solr -i siard-1 -if c:\software\database\mestieri_ASCS.siard
```

Scenari

Conservazione database

Database

Conservazione contenuti

Conservazione web

Scenario 2: ASMM

L'Archivio Storico Multimediale del Mediterraneo (ASMM) è un sistema informativo archivistico sviluppato nel 2006 dedicato alla gestione di documenti e raccolte cartografiche conservate negli Archivi di Stato relative al Mar Mediterraneo.

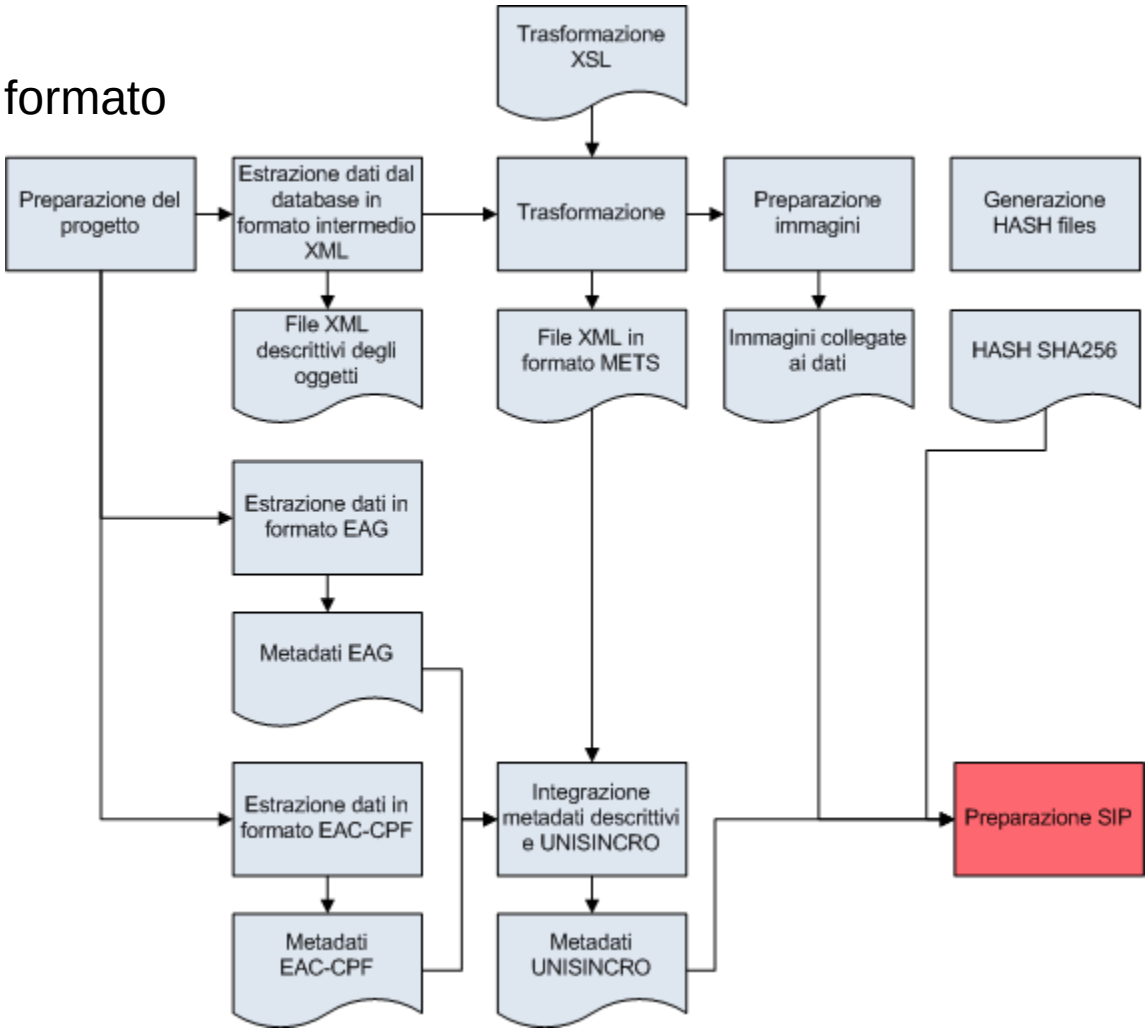
La banca dati contiene 62.000 schede descrittive e 350.000 immagini a media e bassa risoluzione per una occupazione di circa un TB

L'infrastruttura si basa su di un RDBMS Oracle 10g ed è divenuto obsoleto in termini di hardware e software.

La struttura originale del database era estremamente complessa e notevolmente ridondante rispetto alle esigenze descrittive archivistiche: 28 tabelle nel database principale, 42 nel database di indicizzazione e 30 nel content management system.

Processo

- ✓ Estrazione dati dal database in formato intermedio XML
- ✓ Trasformazione dei contenuti
- ✓ Estrazione dati in formato EAG
- ✓ Estrazione dati in formato EAC
- ✓ Preparazione immagini
- ✓ Generazione HASH files
- ✓ Integrazione metadati descrittivi e Unisincro



Estrazione dei contenuti

Si sono estrapolati i contenuti del database e si sono strutturati utilizzando i metadati: EAD,EAC-CPF, EAG, METS SAN..

I contenuti del database sono stati esportati in un metaformato xml per sganciarsi dalla gestione del motore del database.

Si è partiti dall'analisi della struttura della base di dati, utilizzando la documentazione di analisi iniziale per elaborare uno schema dei campi utilizzati e per effettuare una mappatura fra i campi originali e quelli degli standard utilizzati.

Tutti i dati di interesse archivistico sono quindi disponibili per poter essere elaborati e/o trasformati e/o conservati.

Le informazioni estratte dal database sono state organizzate gerarchicamente a partire dal soggetto conservatore fino ai livelli di unità documentaria e cartografiche ciascuna delle quali contiene i collegamenti alle immagini.

Estrazione dei contenuti

Portale ASMM

Database Oracle ASMM

Procedura di estrazione

XML

Contenuti in metalinguaggio XML

Procedura di conversione

XML Dati in formato EAG

XML Dati in formato EAD

XML Dati in formato EAC-CPF

XML Dati in formato METS-SAN

Generazione meta dati UNISINCRO

XML UNISINCRO

Generazione metadati PREMIS

XML Metadati PREMIS

Estrazione dei contenuti

Le informazioni estratte dal database sono state organizzate gerarchicamente a partire dal soggetto conservatore, con directory di livello inferiore che contengono almeno un file xml che descrive o un fondo o una serie o una sottoserie, fino ad arrivare alle directory delle unità archivistiche, documentarie o cartografiche ciascuna delle quali contiene le immagini.

Per ogni entità archivistica autonoma sono presenti dei metadati in formato xml con eventualmente il collegamento alle immagini, se esistono e se coerenti con il livello descrittivo.

I file xml relativi alle unità (Archivistiche, Documentali, Cartografiche) hanno al loro interno un'area specifica dove è indicato il path relativo dell'immagine collegata.

Estrazione dei contenuti

- IT-ASBA (directory)
- IT-ASBA.xml Archivio di Stato di BARI
  - F9200038 (directory)
  - F9200038.xml Fondo Archivi Notarili
    - S9200048 (directory)
    - S9200048.xml Serie Notai di Bitonto
      - SS02
      - SS02.xml Sottoserie Notaio Angelo de Bitritto
        - UA01
        - UA01.xml Unità Archivistica Protocollo notarile (aa. 1458-86)

```
<IMAGES300>
<image>IT-ASBA/F9200038/S9200048/SS02/UA01/300DPI/0001.jpg</image>
<image>IT-ASBA/F9200038/S9200048/SS02/UA01/300DPI/0226.jpg</image>
<image>IT-ASBA/F9200038/S9200048/SS02/UA01/300DPI/0002.jpg</image>
<image>IT-ASBA/F9200038/S9200048/SS02/UA01/300DPI/0003.jpg</image>
<image>IT-ASBA/F9200038/S9200048/SS02/UA01/300DPI/0085.jpg</image>
<image>IT-ASBA/F9200038/S9200048/SS02/UA01/300DPI/0225.jpg</image>
<image>IT-ASBA/F9200038/S9200048/SS02/UA01/300DPI/0068.jpg</image>
</IMAGES300>
```

Estrazione dei contenuti

```
<?xml version="1.0" encoding="UTF-8"?>
<ISTITUTO>
        <ID>121</ID>
        <ID_PARENT>11</ID_PARENT>
        <TIPOLOGIA>Archivio</TIPOLOGIA>
        <NCTA>ASBA</NCTA>
        <PVCS>Italia</PVCS>
        <PVCR>Puglia</PVCR>
        <PVCC>Bari</PVCC>
        <PVCL>Bari</PVCL>
        <LDCE>Archivio di Stato di BARI</LDCE>
        <PVCI>Via Pietro Oreste, 45</PVCI>
        <PVCP>70123</PVCP>
        <PVCT>080/099311</PVCT>
        <PVCF>080/099311</PVCF>
        <PVCM>ASBA@archivi.beniculturali.it</PVCM>
        <PVCA>http://archivi.beniculturali.it/ASBA</PVCA>
        <ISTD>GIUSEPPE DIBENDETTO</ISTD>

</ISTITUTO>
```

Archivio di Stato di BARI

Estrazione dei contenuti

```
<?xml version="1.0" encoding="UTF-8"?>
<BASEDOCUMENTALE>
        <ID>229524</ID>
        <ID_SOGP>146300</ID_SOGP>
        <NCT>F9200038</NCT>
        <DEN>Archivi Notarili</DEN>
        <DEST>Cartaceo</DEST>
        <ERMS>XV</ERMS>
        <ERMC>1444</ERMC>
        <ERMP>Metà (da 40 a 60)</ERMP>
        <ERES>XIX</ERES>
        <EREC>1893</EREC>
        <EREP>Fine (da 90 a 99)</EREP>
        <STCM>Buono</STCM>
        <CONS>33239, di cui 32496 voll., 543 rep. e 200 indici</CONS>
        <STRC>Elenco cronologico dei Notai del Distretto di Bari per Piazze ordinate alfabeticamente, Indice alfabetico
        onomastico dei Notai del Distretto di Bari, Indice cronologico onomastico dei Notai del Distretto di Bari, Indice dei capitoli
        matrimoniali dei Notai del Distretto di Bari, Indice dei testamenti dei Notai del Distretto di Bari, Inventario topografico, Indice
        dei volumi notarili microfilmati</STRC>
        <DATA_CREAZIONE>2007-02-21 12:55:06.765</DATA_CREAZIONE>
        <ULTIMA_MODIFICA>2008-10-27 15:04:44.687</ULTIMA_MODIFICA>
        <CUID>utcoord</CUID>
        <MUID>utcoord</MUID>
        <IMAGES72> </IMAGES72>
        <IMAGES300> </IMAGES300>
</BASEDOCUMENTALE>
```

Fondo Archivi notarili (F9200038)

Estrazione dei contenuti

```
<?xml version="1.0" encoding="UTF-8"?>
-<BASEDOCUMENTALE>
        <ID>229526</ID>
        <ID_SOGP>146300</ID_SOGP>
        <NCT>S9200048</NCT>
        <DEN>Notai di Bitonto </DEN>
        <DEST>Cartaceo</DEST>
        <ERMS>XV</ERMS>
        <ERMC>1444</ERMC>
        <ERMP>Met&#224; (da 40 a 60)</ERMP>
        <ERES>XIX</ERES>
        <EREC>1893</EREC>
        <EREP>Fine (da 90 a 99)</EREP>
        <STCM>Buono</STCM>
        <CONS>4159, di cui 4103 voll., 40 rep. e 16 indici </CONS>
        <STRC>Elenco cronologico dei Notai del Distretto di Bari per Piazze ordinate alfabeticamente, Indice alfabetico
        onomastico dei Notai del Distretto di Bari, Indice cronologico onomastico dei Notai del Distretto di Bari, Indice dei capitoli
        matrimoniali dei Notai del Distretto di Bari, Indice dei testamenti dei Notai del Distretto di Bari, Inventario topografico,
        Indice dei volumi notarili microfilmati</STRC>
        <DENN>Il soggetto produttore Archivio notarile distrettuale di Bari, per la serie, &#232; responsabile del versamento.
        In realt&#224; i soggetti produttori della serie dovrebbero essere i Notai di Bitonto che hanno prodotto gli
        atti.</DENN>
        <DATA_CREAZIONE>2007-02-21 12:59:44.765</DATA_CREAZIONE>
        <ULTIMA_MODIFICA>2008-10-27 15:05:49.453</ULTIMA_MODIFICA>
        <CUID>utcoord</CUID>
        <MUID>utcoord</MUID>
        <IMAGES72>
        </IMAGES72>
        <IMAGES300>
        </IMAGES300>
</BASEDOCUMENTALE>
```

Serie Notai di Bitonto (S9200048)

ICAR
ISTITUTO CENTRALE PER GLI ARCHIVI

Estrazione dei contenuti


```
<?xml version="1.0" encoding="UTF-8"?>
<BASEDOCUMENTALE>
        <ID>229528</ID>
        <ID_SOGP>146307</ID_SOGP>
        <NCT>SS02</NCT>
        <DEN>Notaio Angelo de Bitritto</DEN>
        <DEST>Cartaceo</DEST>
        <ERMS>XV</ERMS>
        <ERMC>1458</ERMC>
        <ERMP>Terzo quarto (da 50 a 74)</ERMP>
        <ERES>XV</ERES>
        <EREC>1489 </EREC>
        <EREP>Ultimo quarto (da 75 a 99)</EREP>
        <STCM>Discreto</STCM>
        <CONS>6 volumi</CONS>
        <STRC>Elenco cronologico dei Notai del Distretto di Bari per Piazze ordinate alfabeticamente, Indice alfabetico
        onomastico dei Notai del Distretto di Bari, Indice cronologico onomastico dei Notai del Distretto di Bari, Inventario
        topografico, Indice dei volumi notarili microfilmati</STRC>
        <DENN>Le unit&#224; archivistiche hanno subito un intervento di restauro irreversibile all&#191;inizio degli anni Settanta
        presso il  #191;Laboratorio di Restauro del Libro&#191; dell&#191;abbazia della Madonna della Scala di Noci (BA) e
        successivamente sono state nuovamente rilegate per ripristinare l&#191;ordine cronologico originario, probabilmente dopo
        la pubblicazione del volume: Archivio di Stato di Bari, La presenza ebraica in Puglia. Fonti documentarie e bibliografiche,
        Bari [1981], realizzato in occasione del I Convegno internazionale &#191;Italia judaica&#191; organizzato dall&#191;allora
        Ministero per i Beni Culturali e Ambientali. Si &#232; trattato evidentemente di un&#191;operazione complessa che ha
        interessato tutti i volumi prodotti dal notaio Angelo deBitritto (2/43; 2/44; 2/46; 2/47; 2/48), eccetto il volume 2/45
        contrassegnato dall&#191;antico numero di versamento &#191;65&#191;,relativo all&#191;anno 1467, l&#191;unico che
        conserva la legatura del restauro di Noci.A seguito del suddetto intervento, i volumi non si configurano pi&#249; come al
        momento del versamento da parte dell&#191;Archivio notarile di Bari avvenuto nel 1945 (ovvero 4 volumi contrassegnati
        rispettivamente dai numeri: [64] relativo agli anni [1463-1487]; &#191;65&#191; relativo all&#191;anno &#191;1468&#191;;
        [66] relativo all&#191;anno [1477]; [78] relativo all&#191;anno [1469]), ma nell&#191;attuale sequenza di sei volumi,
        contrassegnati dalle segnature archivistiche 2/43; 2/44; 2/45; 2/46; 2/47; 2/48. Tale ricostruzione &#232; stata possibile
        grazie al raffronto tra le diverse numerazioni presenti sui volumi, l&#191;elenco di versamento dell&#191;Archivio notarile
        di Bari datato 10 agosto 1945, e i microfilm di sicurezza realizzati nel 1963, a cura dell&#191;Archivio di Stato di Bari,
        prima dell&#191;intervento di restauro presso l&#191;abbazia di Noci.</DENN>
        <DATA_CREAZIONE>2007-02-21 13:02:12.906</DATA_CREAZIONE>
        <ULTIMA_MODIFICA>2009-07-09 17:35:21</ULTIMA_MODIFICA>
        <CUID>utcoord</CUID>
        <MUID>utcoord</MUID>
        <SOGI>
                <ID>948061</ID>
                <ID_SOGGETTO>146307</ID_SOGGETTO>
                <SOGI>Donni Benedicti de Bitricto, Angelus</SOGI>
        </SOGI>
        <SOGI>
                <ID>1292864</ID>
                <ID_SOGGETTO>146307</ID_SOGGETTO>
                <SOGI>Bitritto, Angelus de</SOGI>
        </SOGI>
        <LOCS>
                <ID>948063</ID>
                <ID_SOGGETTO>146307</ID_SOGGETTO>
                <LOCS>Bitonto [Bari]</LOCS>
        </LOCS>
        <IMAGES72>
        </IMAGES72>
        <IMAGES300>
        </IMAGES300>
</BASEDOCUMENTALE>
```


Sottoserie Notaio Angelo de Bitritto (SS02)

ICAR
ISTITUTO CENTRALE PER GLI ARCHIVI

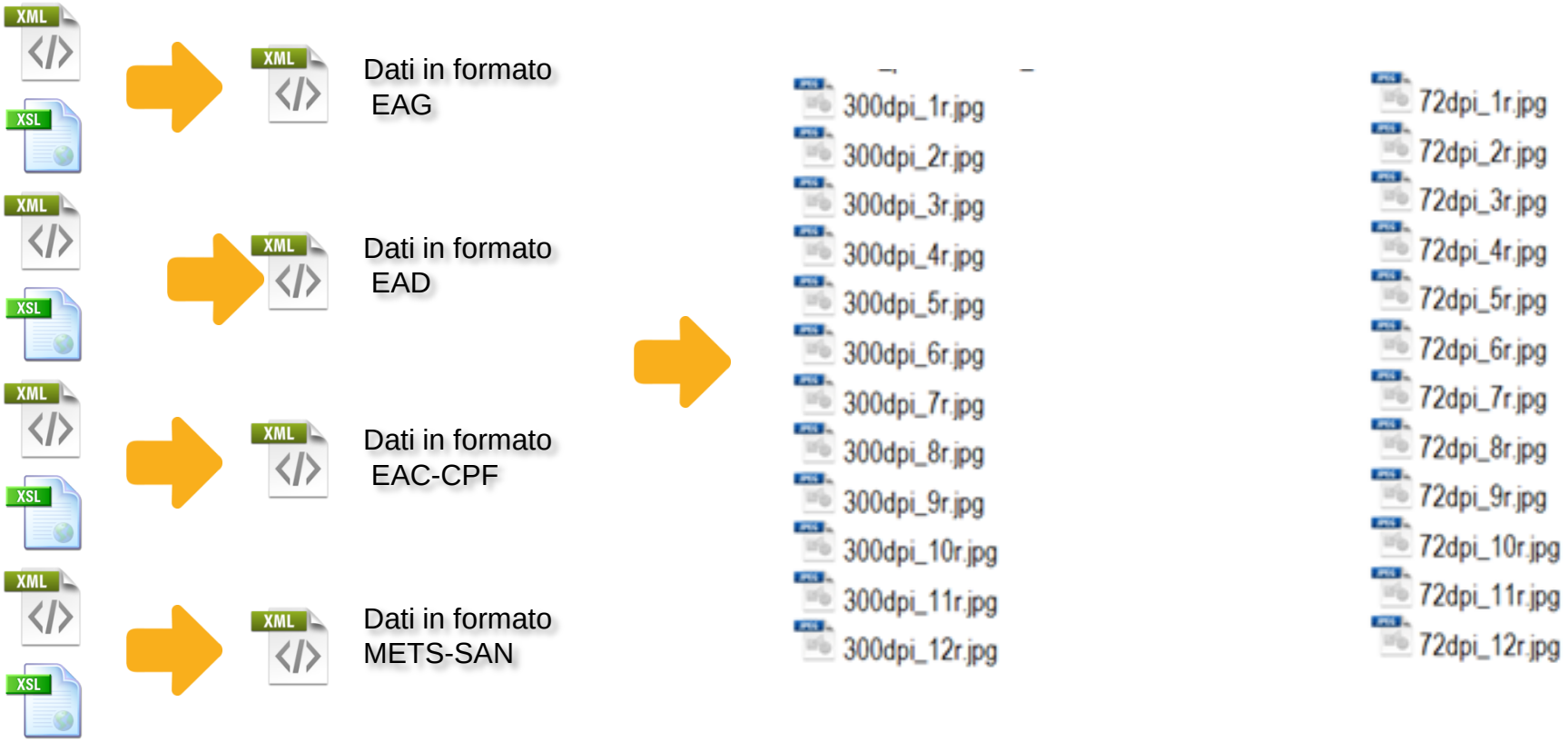
Estrazione dei contenuti


```
<BASEDOCUMENTALE>
    <ID>229530</ID>
    <ID_SOGP>146307</ID_SOGP>
    <NCT>UA01</NCT>
    <DEN>Protocollo notarile (aa. 1458-86)</DEN>
    <DEST>Cartaceo</DEST>
    <ERMS>XV</ERMS>
    <ERMC>&lt;1458&gt; settembre 1</ERMC>
    <ERMV>Data attribuita</ERMV>
    <ERMP>Seconda metà (da 50 a 99)</ERMP>
    <ERES>XV</ERES>
    <EREC>1486 dicembre 22</EREC>
    <EREP>Seconda metà (da 50 a 99)</EREP>
    <CRON>1459 e 1487 secondo l&apos;uso di Bitonto.</CRON>
    <STCM>Discreto</STCM>
    <CONS>112 cc.</CONS>
    <DEAU>Consultazione in originale</DEAU>
    <DENN>Essendo costituito da due parti, il volume presenta una doppia cartulazione (da 1r a 33v; da 1r a 78r). Esso
    contiene nella prima parte atti rogati nell'anno 1486 e nella seconda parte atti rogati negli anni 1458-59. La seconda parte
    del volume, priva di intitolazione e risalente alla Settima Indizione, già datata 1489, in base alla lettura degli atti è stata
    attribuita al 1458-1459 per la presenza di personaggi attivi in quegli anni e di alcuni elementi cronologici (cfr. per esempio il
    doc. 2/43, c. 75r-75v). Le immagini 0001 e 0226 corrispondono alla coperta; l&apos;immagine 0002 è l&apos;intestazione
    della prima parte del volume; le immagini 0003, 0085 e 0225 sono carte bianche; l&apos;immagine 0068 è solo cartulata.
    </DENN>
    <DATA_CREAZIONE>2007-04-03 12:46:58.149</DATA_CREAZIONE>
    <ULTIMA_MODIFICA>2007-10-19 11:41:32.171</ULTIMA_MODIFICA>
    <CUID>utcoord</CUID>
    <MUID>utcoord</MUID>
    <DATA_CREAZIONE>2007-04-03 12:46:58.149</DATA_CREAZIONE>
    <ULTIMA_MODIFICA>2007-10-19 11:41:32.171</ULTIMA_MODIFICA>
    <CUID>utcoord</CUID>
    <MUID>utcoord</MUID>
    <SOGI>
        <ID>948061</ID>
        <ID_SOGGETTO>146307</ID_SOGGETTO>
        <SOGI>Donni Benedicti de Bitricto, Angelus</SOGI>
    </SOGI>
    <SOGI>
        <ID>1292864</ID>
        <ID_SOGGETTO>146307</ID_SOGGETTO>
        <SOGI>Bitritto, Angelus de</SOGI>
    </SOGI>
    <LOCS>
        <ID>948063</ID>
        <ID_SOGGETTO>146307</ID_SOGGETTO>
    <LOCS>Bitonto [Bari]</LOCS>
    </LOCS>
    <IMAGES72>
    <image>IT-ASBA/F9200038/S9200048/SS02/UA01/72DPI/0001.jpg</image>
    <image>IT-ASBA/F9200038/S9200048/SS02/UA01/72DPI/0226.jpg</image>
    <image>IT-ASBA/F9200038/S9200048/SS02/UA01/72DPI/0002.jpg</image>
    <image>IT-ASBA/F9200038/S9200048/SS02/UA01/72DPI/0003.jpg</image>
    <image>IT-ASBA/F9200038/S9200048/SS02/UA01/72DPI/0085.jpg</image>
    <image>IT-ASBA/F9200038/S9200048/SS02/UA01/72DPI/0225.jpg</image>
    <image>IT-ASBA/F9200038/S9200048/SS02/UA01/72DPI/0068.jpg</image>
    </IMAGES72>
    <IMAGES300>
    <image>IT-ASBA/F9200038/S9200048/SS02/UA01/300DPI/0001.jpg</image>
    <image>IT-ASBA/F9200038/S9200048/SS02/UA01/300DPI/0226.jpg</image>
    <image>IT-ASBA/F9200038/S9200048/SS02/UA01/300DPI/0002.jpg</image>
    <image>IT-ASBA/F9200038/S9200048/SS02/UA01/300DPI/0003.jpg</image>
    <image>IT-ASBA/F9200038/S9200048/SS02/UA01/300DPI/0085.jpg</image>
    <image>IT-ASBA/F9200038/S9200048/SS02/UA01/300DPI/0225.jpg</image>
    <image>IT-ASBA/F9200038/S9200048/SS02/UA01/300DPI/0068.jpg</image>
    </IMAGES300>
</BASEDOCUMENTALE>
```


Unità Archivistica Protocollo notarile (aa. 1458-86) (UA01)

Trasformazione dei contenuti

La procedura di conversione utilizza i file xml di input e dei file xsl per la trasformazione dei contenuti o per integrare contenuti.



Trasformazione dei contenuti

Le informazioni estrapolate dal database e integrate con gli altri metadati descrittivi sono state inserite in un Indice di Conservazione secondo lo standard Unisincro.

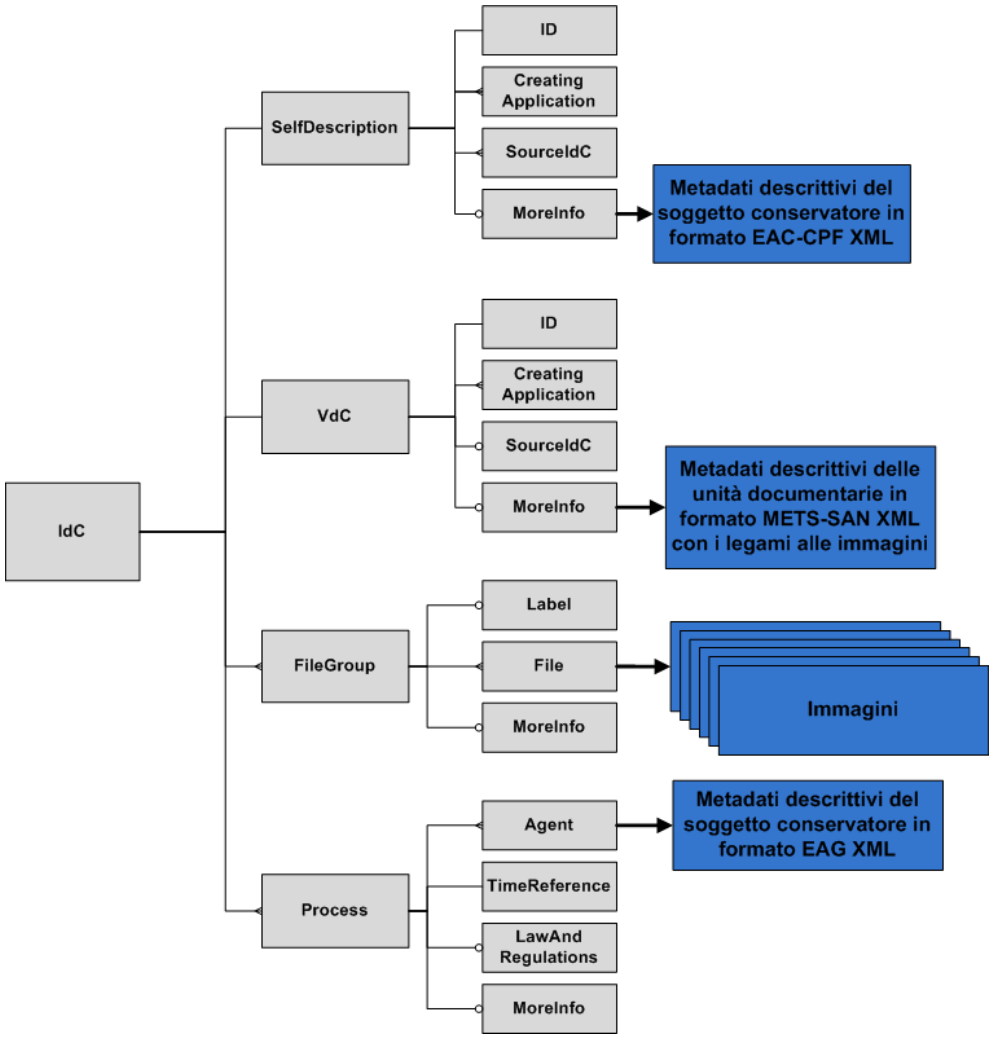
I metadati generati sono poi inseriti in una struttura di metadati Unisincro creando un pacchetto con le immagini.

Sono stati utilizzati i metadati descrittivi in formato EAG per descrivere il soggetto che conserva gli originali analogici,

Sono stati riportati i metadati descrittivi EAC-CPF per descrivere il soggetto produttore .

Le informazioni sulle singole unità documentarie sono state incapsulate in un file METS XML con le descrizioni in formato EAD.

Pacchetto di conservazione per gli oggetti digitali



RODA-IN

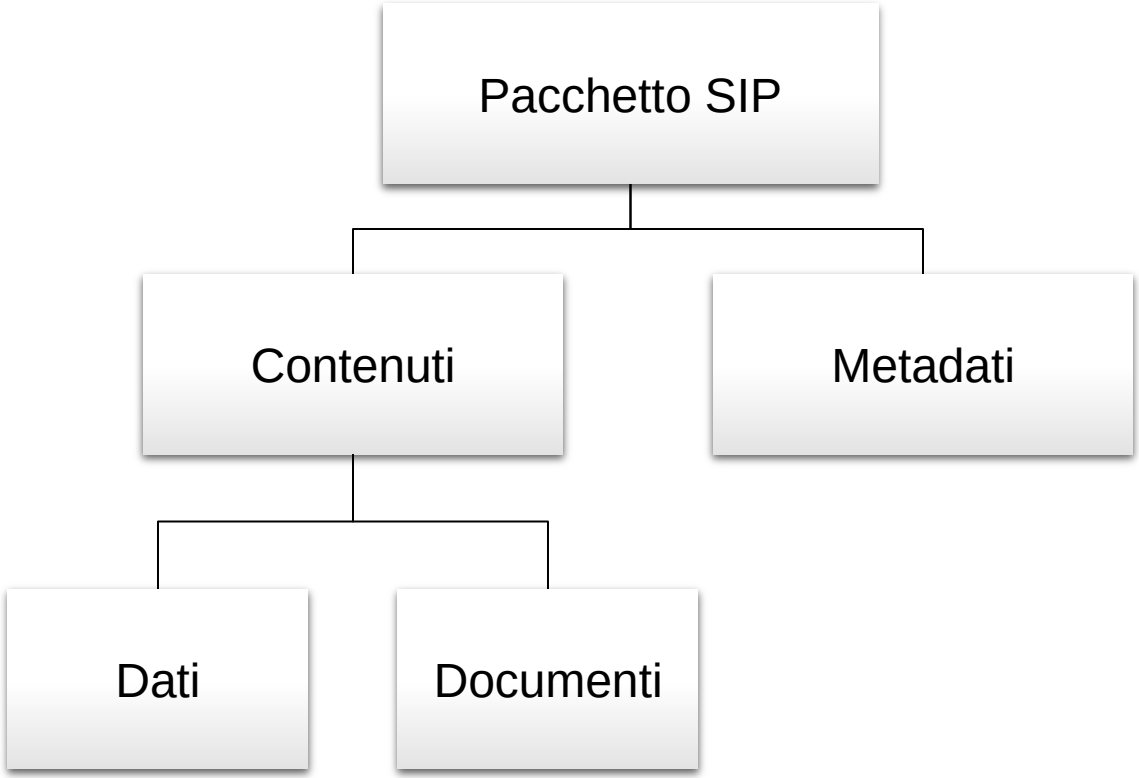
RODA-IN è uno strumento progettato per creare Submission Information Package (SIP) da sottoporre ad un Open Archival Information System (OAIS).

Lo strumento crea SIP da file e cartelle disponibili nel file system locale ed associarli a metadati.

Lo strumento include funzionalità quali:

- Creare, caricare e modificare gli schemi di classificazione/organizzazione
- Associazione automatica di file / cartelle e metadati a SIP
- Definizione di modelli di metadati
- Supporto per vari formati di metadati (EAD, DC, ecc)
- Creazione di SIP di dimensioni illimitate
- Creazione di SIP in vari formati: BagIt e E-ARK

E-Ark SIP: modello concettuale



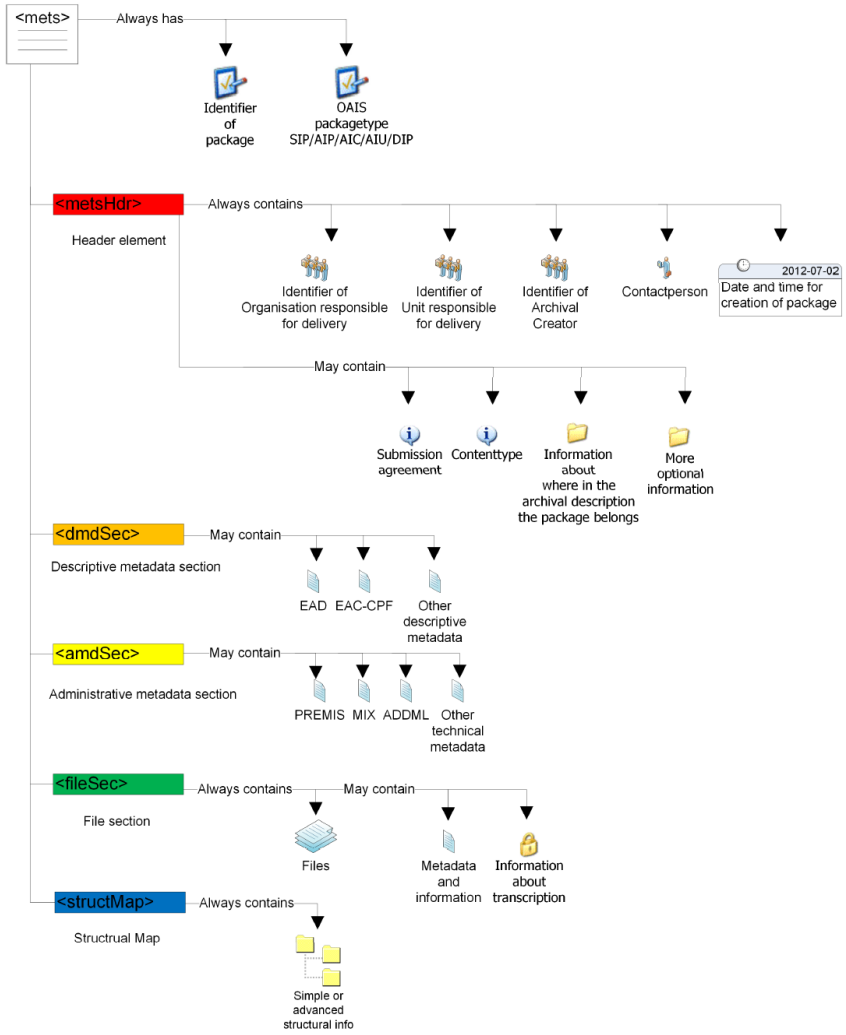
SIP

Elemento Principale

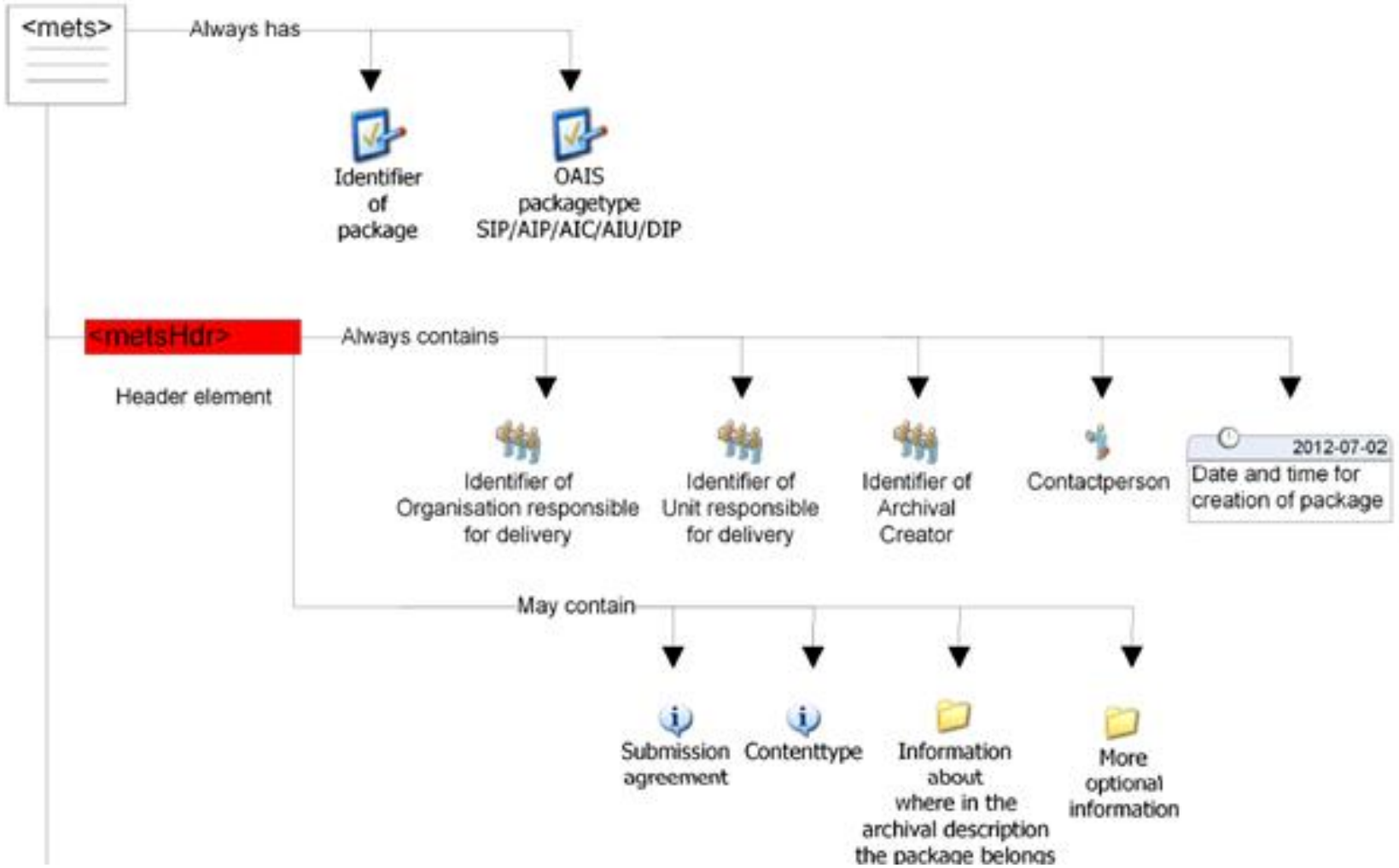
Componente

ISTITUTO CENTRALE PER GLI ARCHIVI

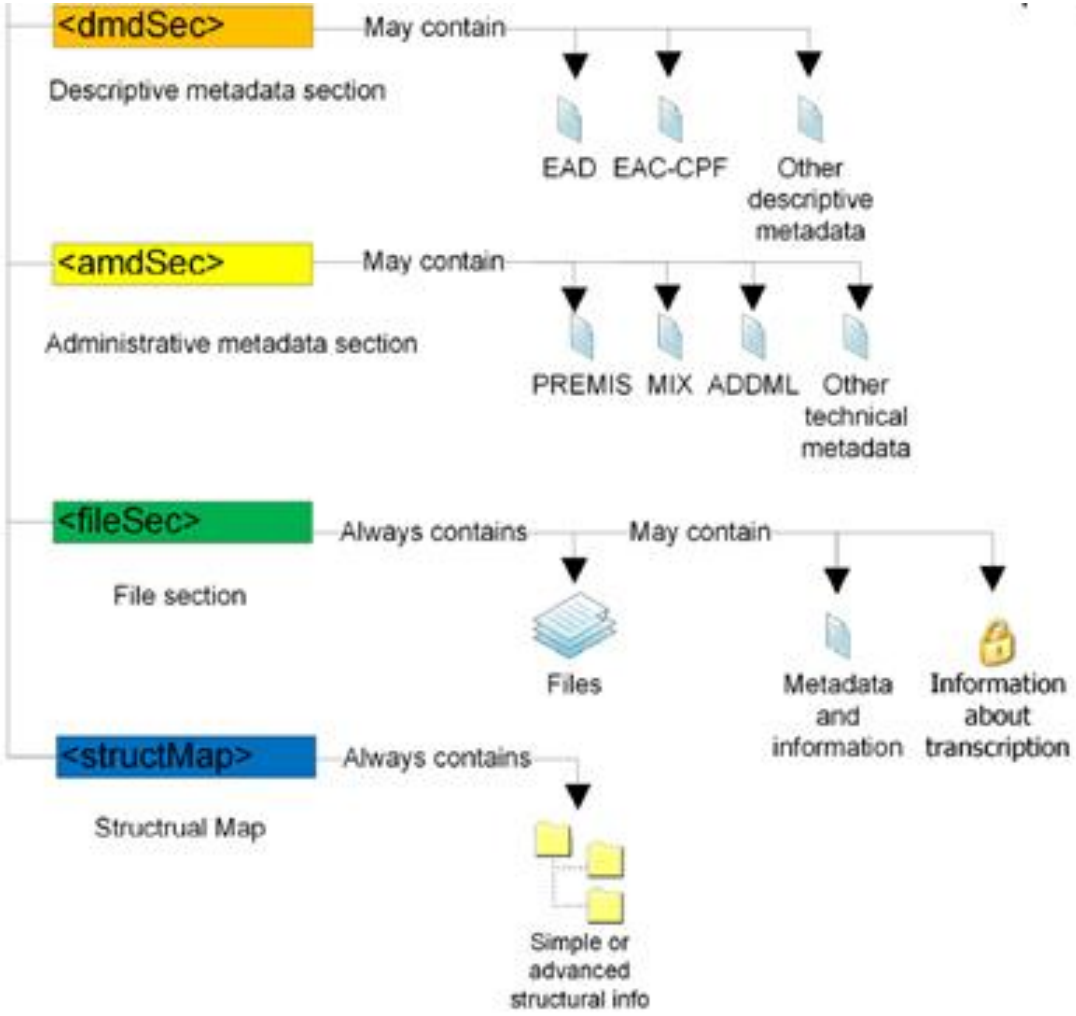
E-Ark SIP: schema METS



E-Ark SIP: schema METS



E-Ark SIP: schema METS



Considerazioni

La conservazione dei database "attivi" che non hanno concluso il loro ciclo di vita, può avere due strategie di approccio con il comune denominatore dell' attenzione ad elementi di contesto e di provenienza. :

- **Estrapolazione dei documenti** informativi rilevanti e loro gestione in autonomia rispetto al database originale

- **Conservazione dell'intero database** secondo i **processi evidenziati** in precedenza

Considerazioni

Nella progettazione dei database e delle applicazioni di gestione documentale (e non solo) bisogna iniziare a tenere in considerazione un approccio che garantisca una corretta conservazione dei contenuti.

- Integrazione e gestione degli eventi e degli agenti legati ai processi (modello del processo, dati di esecuzione del processo, autorizzazioni e schema degli utenti e delle funzioni abilitate su processo), ai documenti (gestione dei legami fra agenti, eventi ed altri oggetti come processi e documenti) e agli utenti/agenti;
- Gestione e monitoraggio delle autenticazioni (Firme digitali, certificazione della titolarità del processo, metadati specifici etc etc)
- Inserimento delle informazioni descrittive ed operative dei relative al/ai processo/i di produzione del documento;

Considerazioni

- Memorizzazione dello stato di ogni singola istanza del/dei processi di produzione (Oggetti, Agenti, Eventi, etc etc)
- Memorizzazione delle intellectual entities e delle loro relazioni per gestire la corretta interpretazione dei documenti che saranno sottoposti a conservazione;
- Memorizzazione delle informazioni relative al processo di aggregazione di campi per costituire un documento informatico. Strutturazione del mapping fra i campi interessati, i contenuti di questi campi ed i campi di arrivo (… logica di aggregazione …) e sua memorizzazione senza possibilità di modifica (… memorizzata in forma statica …). Ad esempio un file xsl o xml specifico di trasformazione firmato digitalmente, memorizzato nel sistema di gestione documentale e riportato nel sistema di conservazione .

Considerazioni

- *Memorizzazione nel database dei dati relativi al contesto archivistico fra cui*
  - *Relazioni con altri documenti (record?) e con altre aggregazioni documentali (recordset?)*
  - *Relazioni con titolari di classificazione (da conservare), registri, elenchi,*
  - *Dati identificativi e descrittivi degli uffici o enti produttori*
  - *Dati identificativi e descrittivi dell'ente conservatore*
  - *Gestione della persistenza delle informazioni esterne (URL, Schemi XSD, ontologie, etc etc)*
- *Gestione dei log del sistema informatico*
  - *Dati di transazione (autenticità, integrità …)*
  - *Strutturazione e Standardizzazione*

WEB Archiving

Il settore dei Beni Culturali ha prodotto e continua a produrre una quantità sempre maggiore di contenuti digitali che hanno bisogno di essere archiviati, conservati e tutelati nel tempo in modo affidabile per consentire che queste risorse possano essere utilizzate in futuro.

Gli aspetti della **conservazione** del **patrimonio digitale** sono stati sottovalutati nella stragrande maggioranza delle iniziative di digitalizzazione del patrimonio e in quelle di costruzione e catalogazione di contenuti culturali.

WEB Archiving

Si assiste sempre di più alla loro scomparsa o impossibilità d'uso concreta con la conseguente perdita della loro valenza culturale e storica e delle risorse umane ed economiche impegnate.

I **siti web** che trattano **contenuti digitali culturali devono** essere sottoposti a **conservazione** e alla **fruizione** attraverso **processi di web archiving.**

Web archiving è il processo di raccolta e di conservazione di siti web allo scopo di creare una loro storicizzazione a scopi di ricerca e consultazione.

WEB Archiving

Esempio da www.archive.org relativo al sito web del Comune di Roma

Novembre 1996

Aprile 2017

1.195 snapshot dal1 novembre 1996 al 12 Marzo 2017

WEB Archiving

WEB Archiving: conservazione di mostre virtuali

Mostre MOVIO dell'Istituto Centrale per gli Archivi

***Terzo scenario*: Conservazione di una mostra virtuale relativa a materiale archivistico:**

- Lavoro completamente digital born
- Uso di una piattaforma di Web Content Management in un sistema in hosting

Problematiche

- Quali formati per la conservazione dei siti web?
- Processi di web crawling
- Quali metadati per la conservazione e la successiva reperibilità dei contenuti?
- Come costruire il pacchetto di conservazione?
- Nessuna pianificazione della conservazione

WEB Archiving: conservazione di mostre virtuali

Il teatro nel fascismo - Rappresentazione e censura nei documenti d'archivio (1931 -1944)

WEB Archiving: conservazione di mostre virtuali

Alla corte degli Albani - Testimonianze di una nobile famiglia bergamasca attraverso il loro Album di Disegni (XIX secolo)

WEB Archiving: conservazione di mostre virtuali

Le tavolette di Biccherna - Storia costume e società: l'immagine della città di Siena attraverso le tavolette di Biccherna

WEB Archiving: conservazione di mostre virtuali

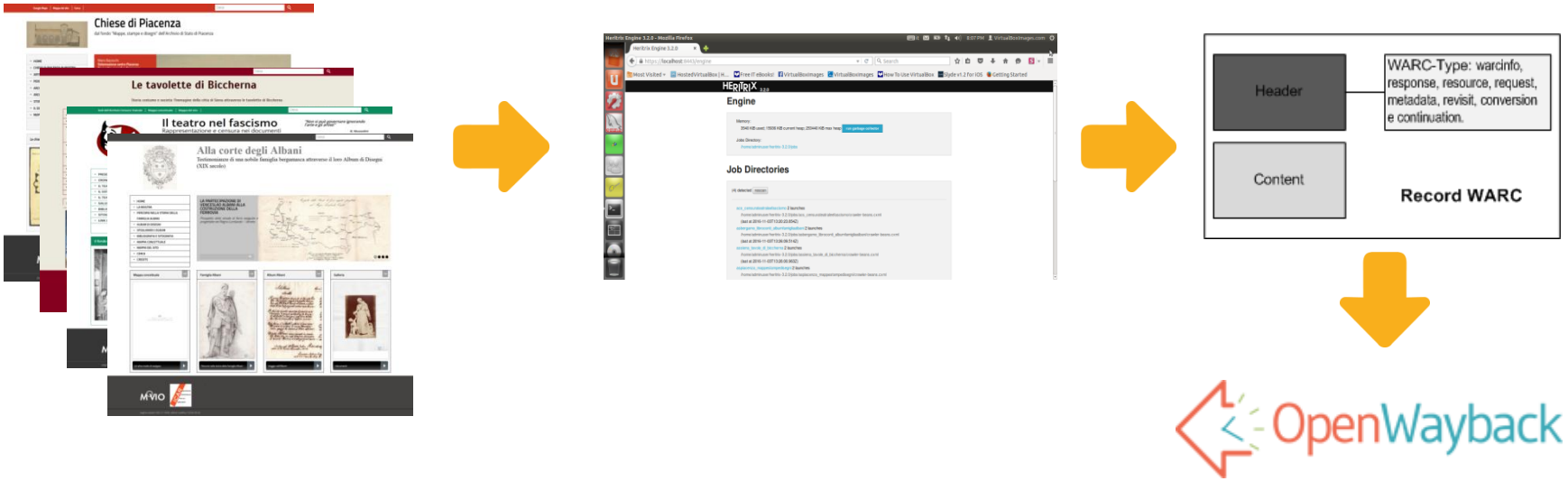
Chiese di Piacenza dal fondo

"Mappe, stampe e disegni"

dell'Archivio di Stato di Piacenza

WEB Archiving: processo

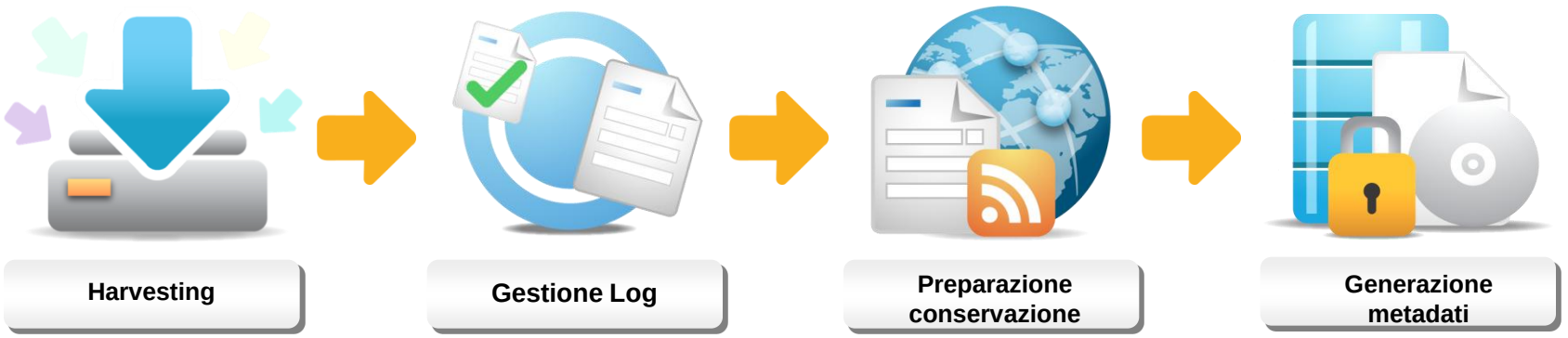
La tecnica del **web archiving** è stata utilizzata in questo lavoro per recuperare quei contenuti disponibili attraverso applicazioni web che interfacciano un database e contenuti digitali.



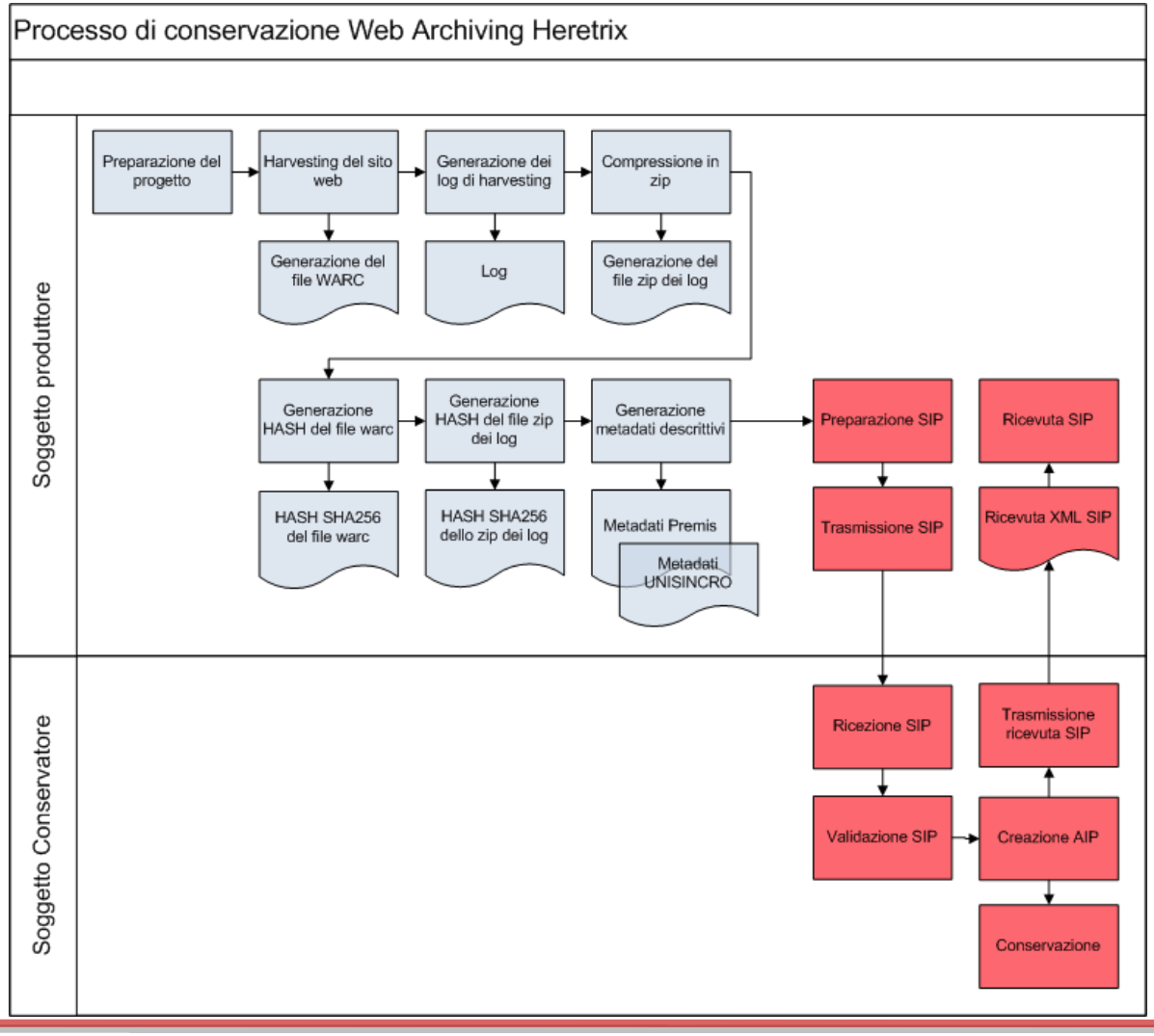
Processo di WEB Archiving

Il processo implementato prevede 4 passaggi:

1. Preparazione ed harvesting del sito web
2. Gestione dei log di harvesting (generazione, compressione, hashing)
3. Preparazione alla conservazione ed hashing del file WARC
4. Generazione metadati descrittivi PREMIS e UNISINCRO



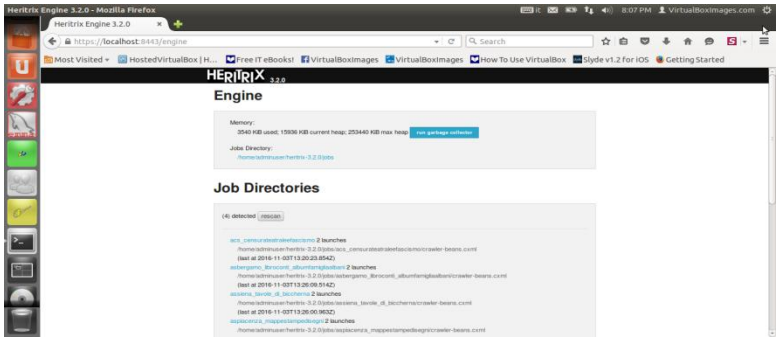
Processo di WEB Archiving



ICAR
ISTITUTO CENTRALE PER GLI ARCHIVI

Harvesting

Pianificazione delle attività



Harvesting con Heritrix

Url Progetto	Dim. WARC	Elementi
http://movio.beniculturali.it/icar/asbergamo_libroconti_albumfamigliaalbani	63.672.276	642
http://movio.beniculturali.it/icar/acs_censurateatraleefascismo	54.785.293	1027
http://movio.beniculturali.it/icar/assiena_tavole_di_biccherna	81.632.291	968
http://movio.beniculturali.it/icar/aspiacenza_mappestampedisegni	82.788.852	1309

Gestione dei log di harvesting

Nome del file di log	Informazioni
Crawl.log	Informazioni sull'esecuzione dell'intero processo di crawling
Nonfatal-errors	Errori non bloccanti in fase di crawling
Uri-errors.log	Errori in fase di crawling
Alerts.log	Segnalazioni generiche
Runtime-errors-log	Errori del sistema di crawling in fase di esecuzione

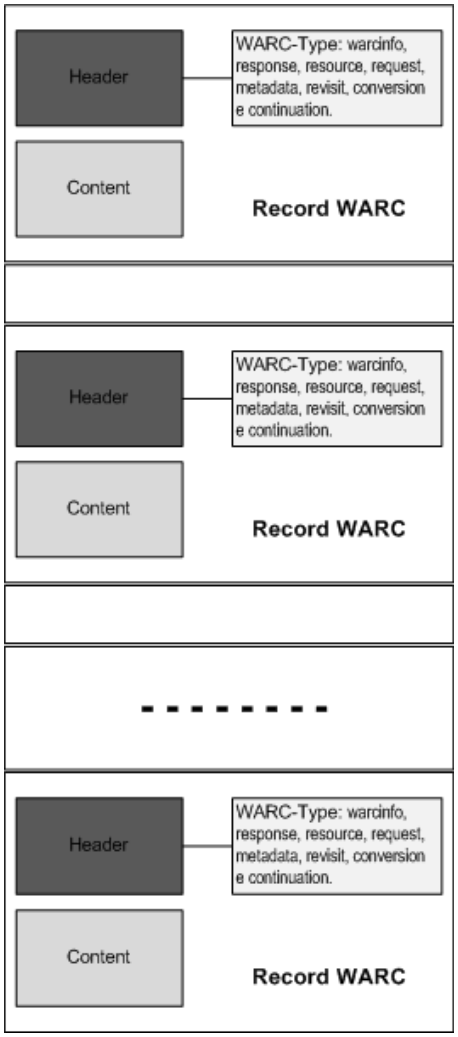
Ogni job del programma heritrix genera una serie di report e di log utili per controllarne il funzionamento e da conservare nel pacchetto di versamento per verificare la qualità e la completezza dell'harvesting.

Compressione del file

Hashing del file

Preparazione WARC

Preparazione conservazione



Hashing del file (SHA256)

5B9F4CA9A10387BFAF0EC884E5822DE9
8AA3470B177C1C7A0368C84D87AB2DEF

Metadati Premis

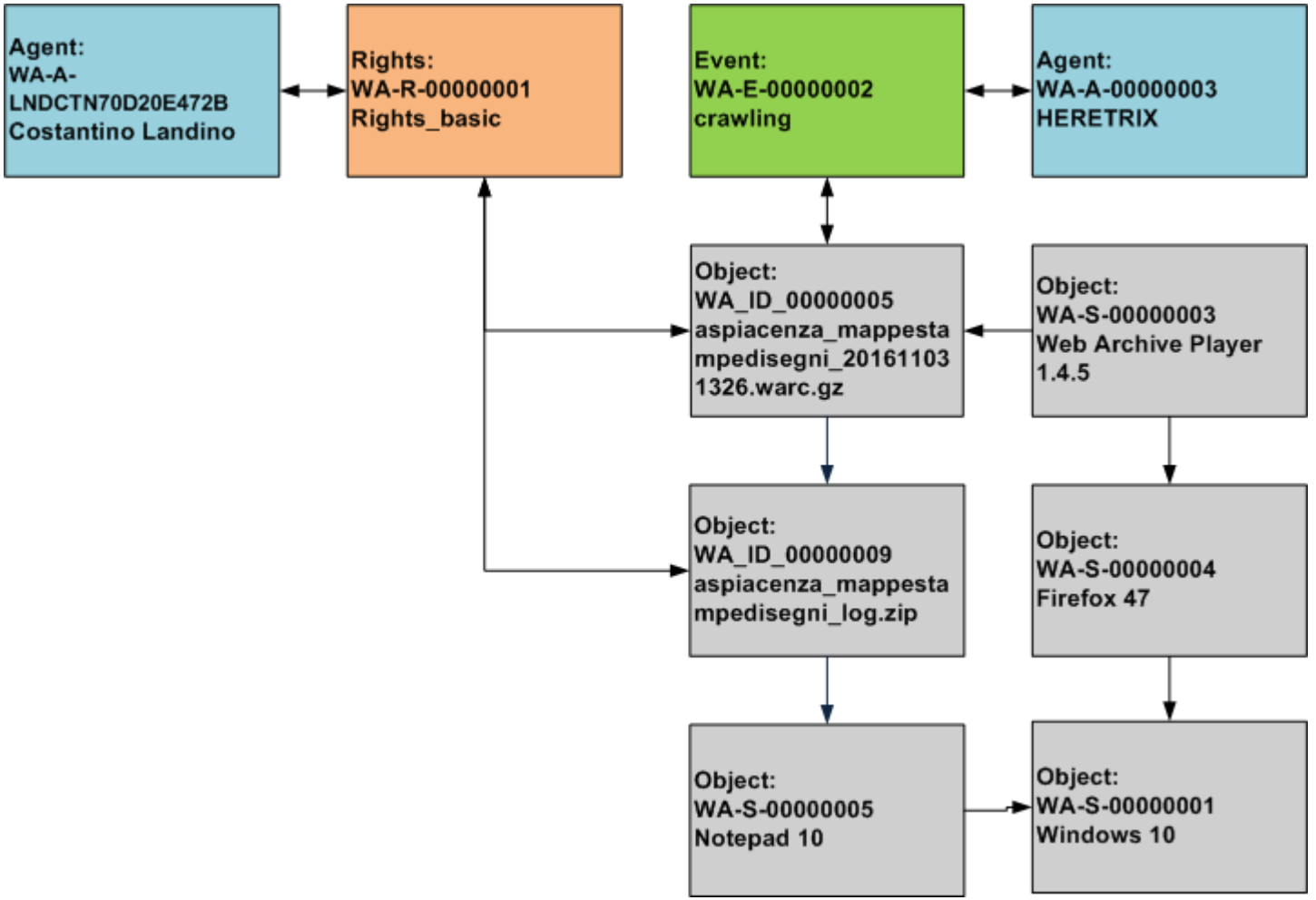
Generazione metadati

Gli oggetti interessati alla conservazione avranno un loro identificativo che sarà utilizzato nei metadati PREMIS ed UniSincro per la gestione delle relazioni fra le varie entità.

Entità	codice identificativo	Contenuti
Object	WA_ID_00000005	aspiacenza_mappestampedisegni_201611031326.warc.gz
Object	WA_ID_00000006	acs_censurateatraleefascismo_201611031321.warc.gz
Object	WA_ID_00000007	asbergamo_libroconti_albumfamigliaalbani_201611031327.warc.gz
Object	WA_ID_00000008	assiena_tavole_di_biccherna_201611031326.warc.gz
Object	WA_ID_00000009	acs_censurateatraleefascismo_logs.zip
Object	WA_ID_00000010	asbergamo_libroconti_albumfamigliaalbani_logs.zip
Object	WA_ID_00000011	aspiacenza_mappestampedisegni_logs.zip
Object	WA_ID_00000012	assiena_tavole_di_biccherna_logs.zip
Object	WA-S-00000001	Windows 10
Object	WA-S-00000003	Web Archive Player 1.4.5
Object	WA-S-00000004	Firefox 47
Object	WA-S-00000005	Notepad 10
Events	WA-E-00000002	Crawling
Agents	WA-A-HERETRIX	HERITRIX
Agents	WA-A-LNDCTN70D20E472B	Costantino Landino
Rights	WA-R-Rights_basic	Diritti generali di accesso

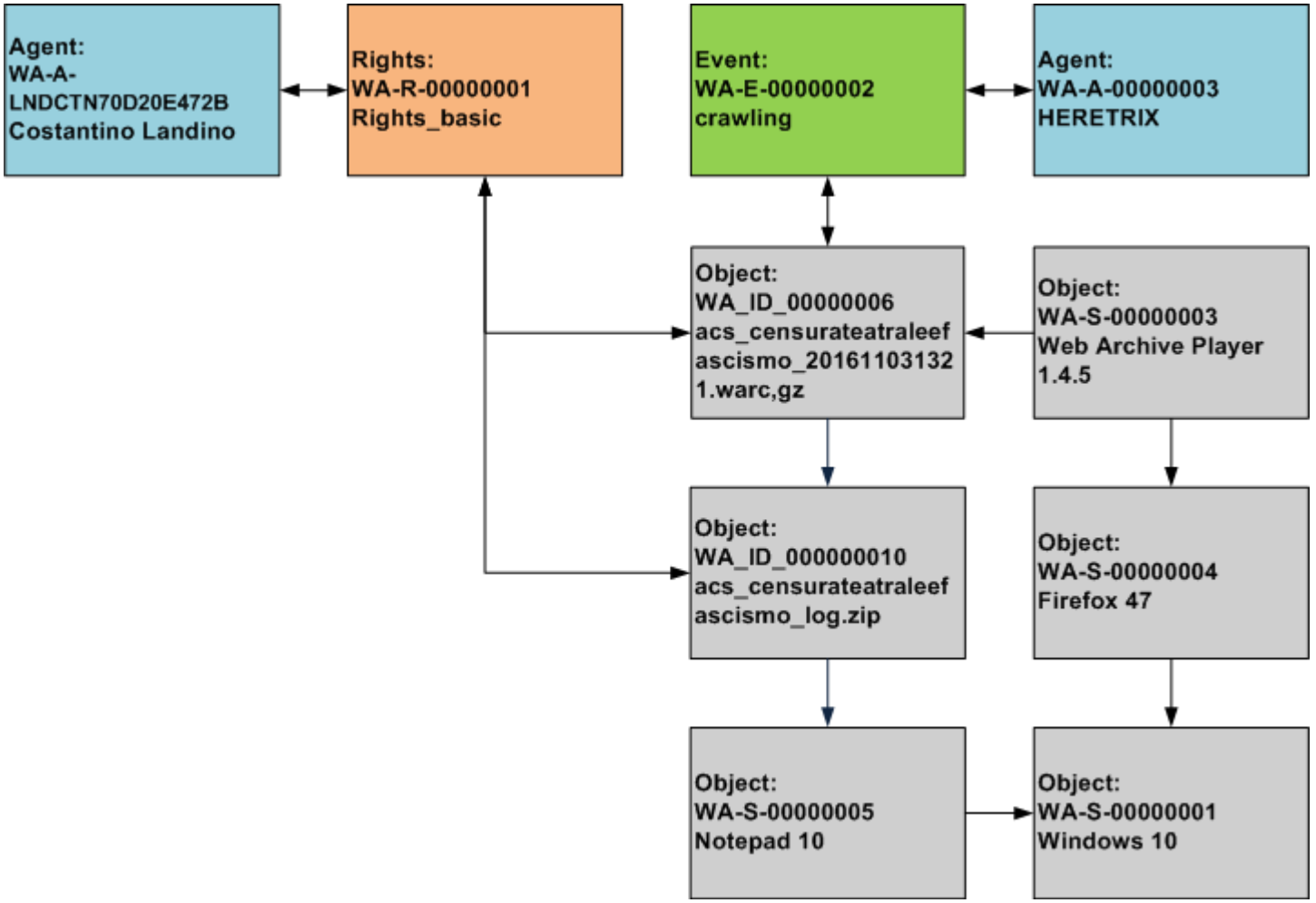
Relazioni nei metadati Premis: Piacenza

Generazione metadati



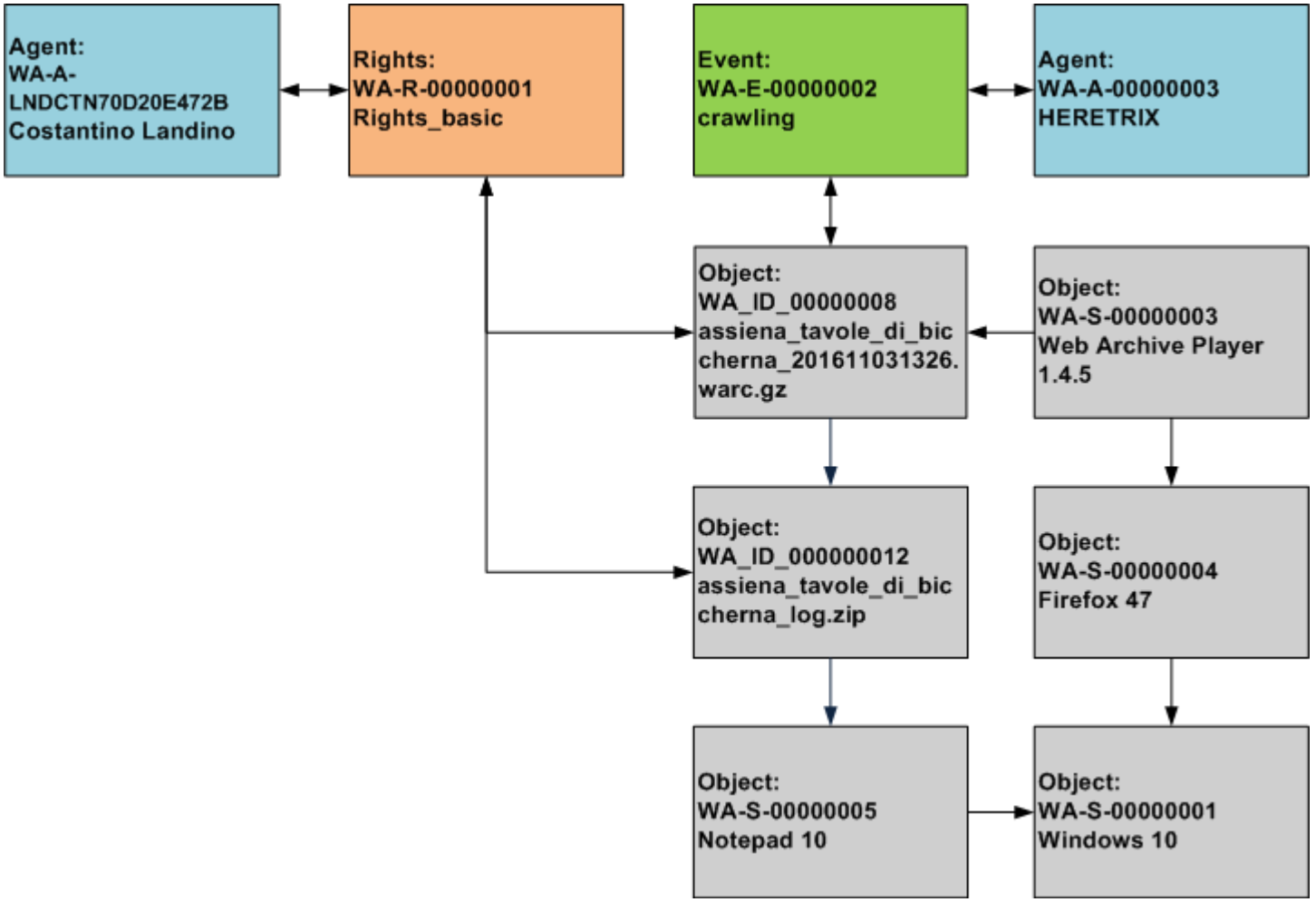
Relazioni nei metadati Premis: Teatro

Generazione metadati



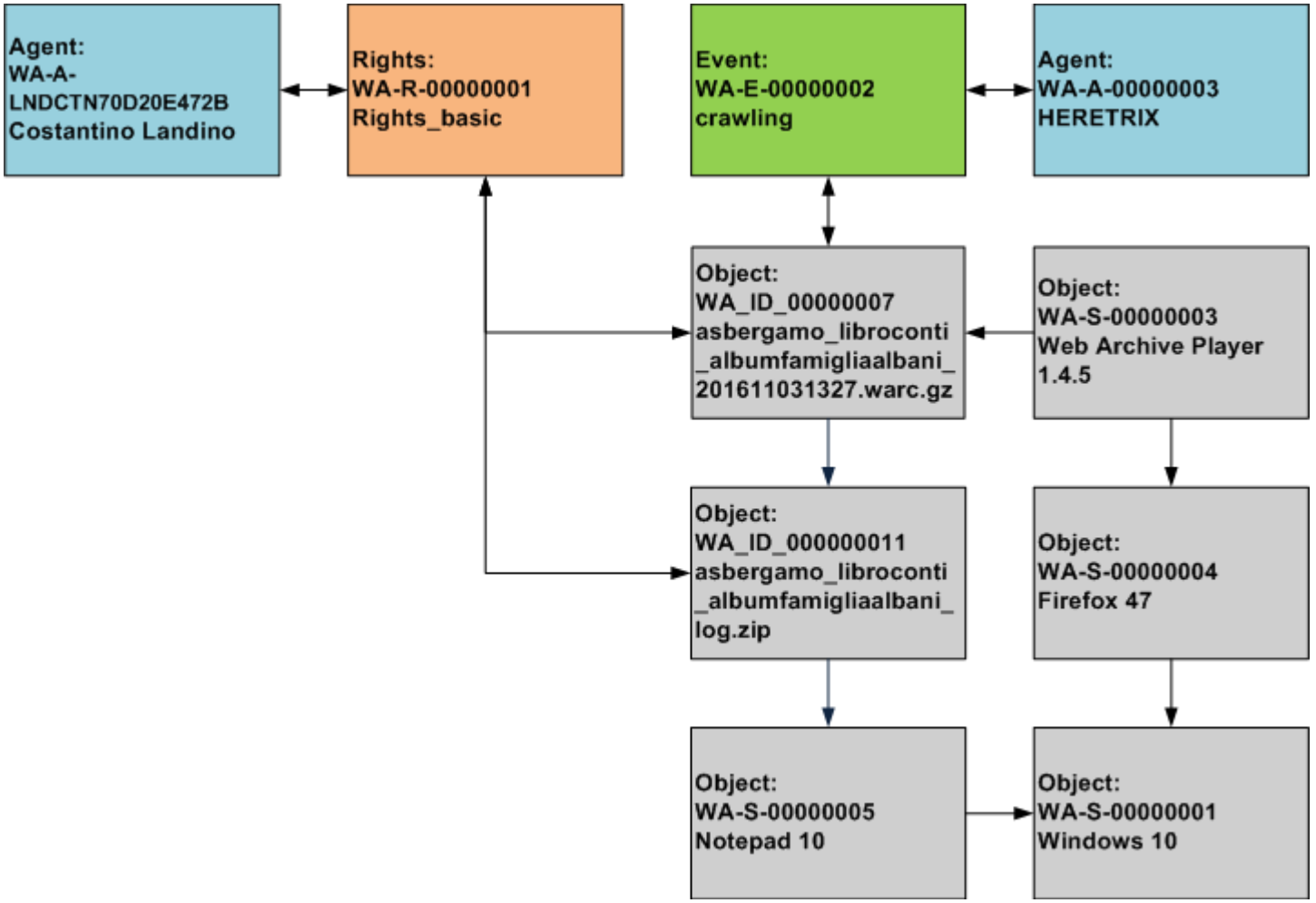
Relazioni nei metadati Premis: Siena

Generazione metadati



Relazioni nei metadati Premis: Bergamo

Generazione metadati



Metadati UNISINCRO

```
<?xml version="1.0" encoding="UTF-8"?>
<sincro:IdC xsi:schemaLocation="unisincro.xsd" xmlns:sincro="http://www.cnipa.gov.it/sincro/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
<sincro:SelfDescription>
        <sincro:ID sincro:scheme="local">WA_ID_00000005</sincro:ID>
        <sincro:CreatingApplication>
        <sincro:Name>Web Conservation  process</sincro:Name><sincro:Version>1.0</sincro:Version>
        <sincro:Producer>Costantino Landino</sincro:Producer>
        </sincro:CreatingApplication>
</sincro:SelfDescription>
<sincro:VdC>
        <sincro:ID sincro:scheme="local">VDC_WA_ID_00000005</sincro:ID>
        <sincro:VdCGroup>
                <sincro:Label>Sito web aspiacenza_mappestampedisegni 20161103</sincro:Label>
                <sincro:ID sincro:scheme="local">VDGG_WA_ID_00000005</sincro:ID>
                <sincro:Description sincro:language="IT"/>
        </sincro:VdCGroup>
        <sincro:MoreInfo sincro:XMLScheme="https://www.loc.gov/standards/premis/premis.xsd">
                <sincro:EmbeddedMetadata>----------/sincro:EmbeddedMetadata>
        </sincro:MoreInfo>
</sincro:VdC>
```

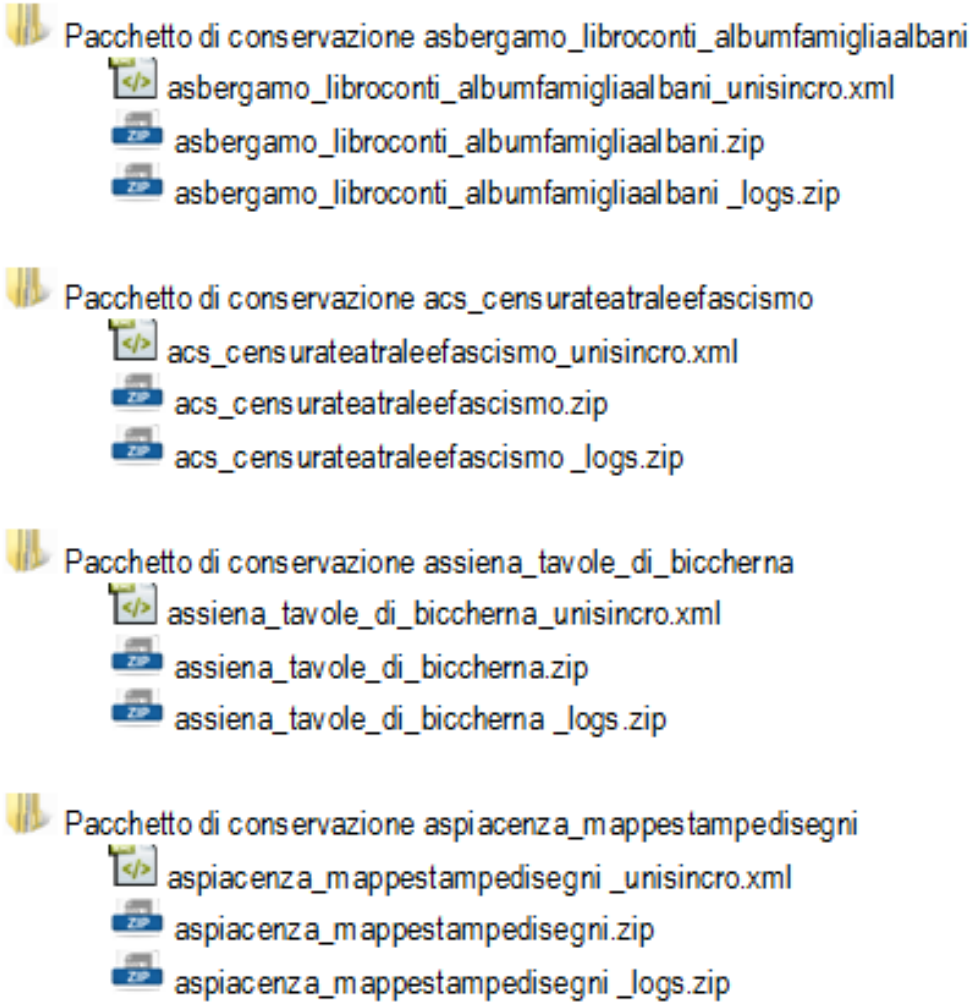
Metadati UNISINCRO

```
<sincro:FileGroup>
        <sincro:Label>aspiacenza_mappestampedisegni_201611031326.warc.gz</sincro:Label>
        <sincro:File sincro:encoding="binary" sincro:format="application/warc">
        <sincro:ID>WA-G-00000005-1</sincro:ID>
        <sincro:Path>aspiacenza_mappestampedisegni_201611031326.warc.gz</sincro:Path>
        <sincro:Hash sincro:function="SHA-256">
80013C26E709CC630748B75FF8427128D9967A1F0330660D39AC4E1319223CB
        </sincro:Hash>
        </sincro:File>
</sincro:FileGroup>
<sincro:FileGroup>
        <sincro:Label>aspiacenza_mappestampedisegni_logs.zip</sincro:Label>
        <sincro:File sincro:encoding="binary" sincro:format="application/zip">
        <sincro:ID>WA-G-00000005-2</sincro:ID>
        <sincro:Path> aspiacenza_mappestampedisegni_201611031326_logs.zip</sincro:Path>
        <sincro:Hash sincro:function="SHA-256">
D9817609E5B857E72A670A4D4B8FF55C12FCC3861BA3866B87C60AF2955D25AD
        </sincro:Hash>
        </sincro:File>
</sincro:FileGroup>
```

Metadati UNISINCRO

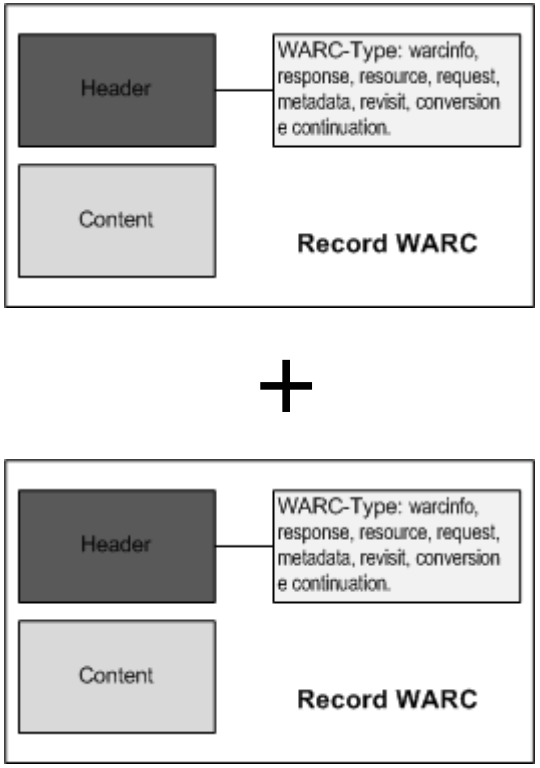
```
<sincro:Process>
        <sincro:Agent sincro:type="person" sincro:role="PreservationManager">
        <sincro:AgentName>
        <sincro:NameAndSurname>
        <sincro:FirstName>Costantino</sincro:FirstName>
        <sincro:LastName>Landino</sincro:LastName>
        </sincro:NameAndSurname>
        </sincro:AgentName>
        <sincro:Agent_ID sincro:scheme="TaxCode">LNDCTN70D20E472B</sincro:Agent_ID>
        </sincro:Agent>
        <sincro:TimeReference>
        <sincro:AttachedTimeStamp sincro:normal="2016-11-18T19:40:00Z"/>
        </sincro:TimeReference>
        </sincro:Process>
</sincro:IdC>
```

Pacchetto di conservazione



Formato WARC

Il formato WARC (Web ARChive) è stato elaborato per la “raccolta” (harvesting), la gestione, l’accesso e lo scambio dei contenuti web.



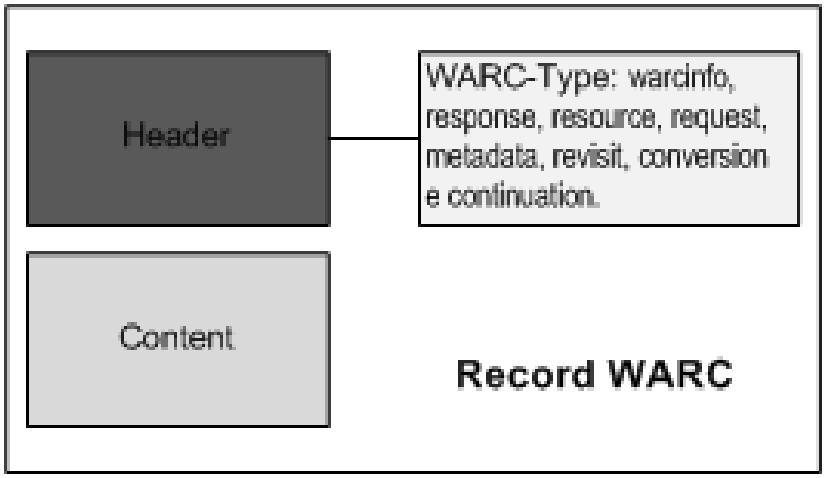
Il formato WARC è di tipo “contenitore” ed è costituito da una sequenza di record/oggetto. Ogni record è composto da un header seguito da un blocco contenuto che costituisce il contenuto vero e proprio.

```
WARC/1.0
WARC-Type: request
WARC-Target-URI: http://movio.beniculturali.it/robots.txt
WARC-Date: 2016-11-03T13:26:29Z
WARC-Concurrent-To: <urn:uuid:4a50c641-cb77-4bde-a334-5085b9956001>
WARC-Record-ID: <urn:uuid:94faa54a-c796-4b29-9346-bac93897e819>
Content-Type: application/http; msgtype=request
Content-Length: 237
GET /robots.txt HTTP/1.0
User-Agent: Mozilla/5.0 (compatible; heritrix/3.2.0 +http://www.costantinolandino.it)
Connection: close
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Host: movio.beniculturali.it
```

Formato WARC

L'Header è costituito da una prima linea che indica la versione del formato (WARC/1.0) seguita da campi del tipo “nome:valore” che servono per fornire varie informazioni sul record (l’URI del sito, data di harvesting, ..); il tutto è concluso con una riga vuota che serve da separatore di blocco. Due righe vuote separano i record.

Un blocco contenuto è costituito dai risultati delle operazioni di harvesting del sito Web (pagine, immagini, redirect, dns request, ..) o da metadati o contenuti trasformati.



Formato WARC

Il formato WARC non ha un meccanismo di compressione ma è possibile utilizzare il metodo GZIP con compressione di tipo “deflate“ che assicura una percentuale di compressione nell'ordine del 60%.

La dimensione media per un file WARC è di 1 GB e, nel caso di dimensioni superiori, è possibile utilizzare più file fra loro correlati.

Il formato WARC è neutro rispetto ai contenuti digitali, permette di memorizzare il flusso delle richieste HTTP, permette di inserire metadati aggiuntivi, gestisce l’assegnazione di un identificativo per ogni file/oggetto, gestisce i duplicati e la segmentazione dei contenuti raccolti su più record quando le dimensioni iniziano ad essere considerevoli.

Formato WARC

La naming convention, proposta da Internet Archive e raccomandata dallo standard ISO; segue lo schema: **Prefix**-**Timestamp**-**Serial**-**Crawlhost**.warc.gz dove

- **Prefix** è l'abbreviazione del progetto;
- **Timestamp** è un timestamp GMT a 14 cifre che indica data e ora di creazione del file;
- **Serial** è un numero seriale possibilmente univoco definito durante il processo di creazione dei file;
- **Crawlhost** è il nome di dominio o l'indirizzo IP della macchina dove è stato creato il file.

Formato WARC

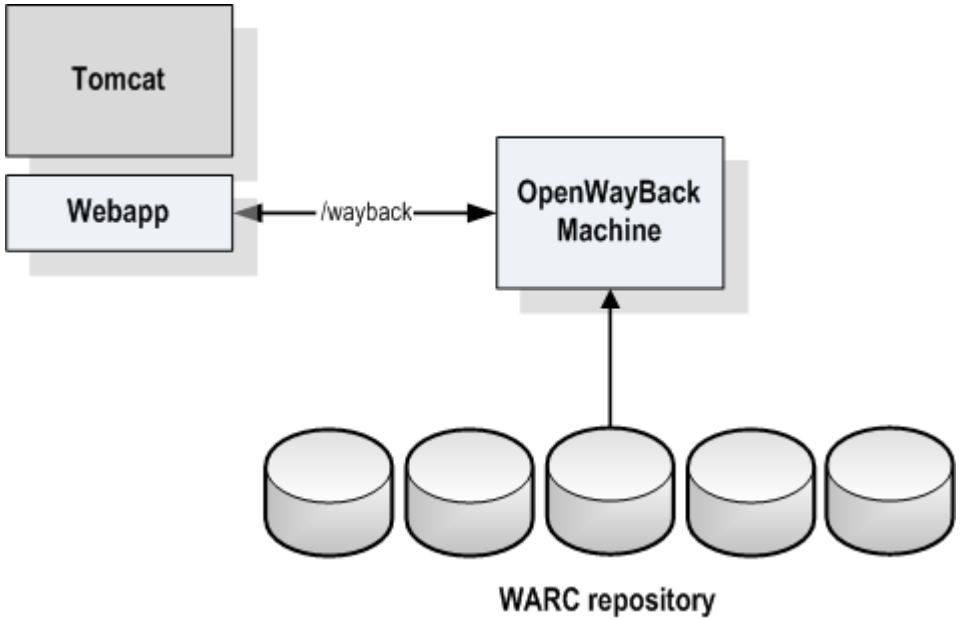
Il formato è adatto e compatibile per un processo di conservazione grazie al rispetto delle caratteristiche di: non proprietà; apertura; standardizzazione e trasparenza.:

- è “non proprietario” in quanto il gruppo di lavoro ISO responsabile del suo mantenimento è il TC46/SC4/WG12;
- è “aperto”, le specifiche del formato sono liberamente disponibili;
- è “standard” in quanto è standard ISO 28500:2009;
- è “trasparente”, in quanto è un formato contenitore per gli oggetti digitali del Web;
- non è sottoposto ad alcuna restrizione (in termini di licenze o brevetti);
- non vi sono meccanismi tecnici di protezione;
- auto-documentato in quanto ciascuna risorsa interna (HTML, JPG, GIF ecc.) ha propri metadati descrittivi.

Waybackmachine

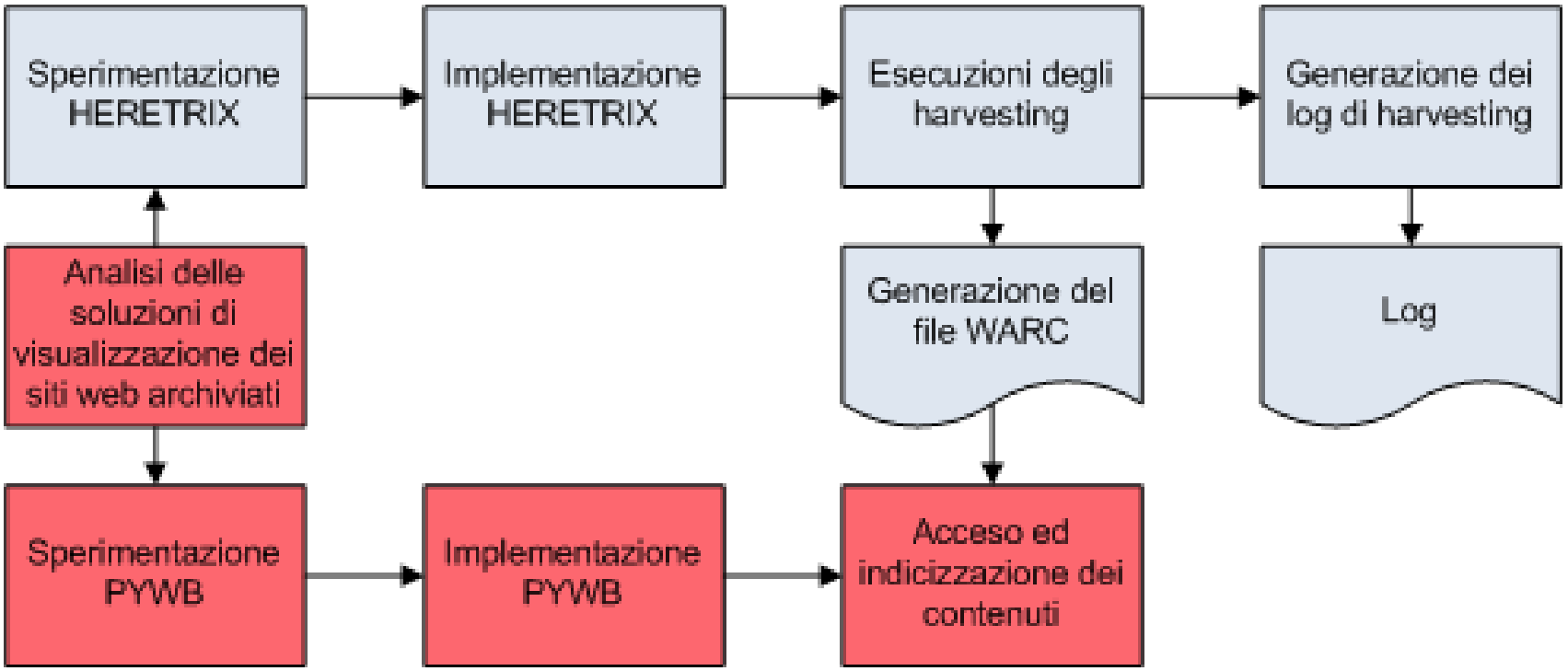
La **Open WayBack Machine** è una completa web application in Java che permette di organizzare e leggere i contenuti dei file .warc e riprodurli come un sito web.

E' utilizzata nell'ambito dell'Internet Archive ed è continuamente aggiornata e migliorata ad opera di una ampia comunità di utenti, nel contesto di iniziative e progetti sparsi in tutto il mondo.



Integrazione nel sito ICAR

Il processo è stato applicato anche al sito ICAR per conservare una copia consultabile via web del sito precedente il completo restyling dello scorso dicembre.



Integrazione nel sito ICAR

Sono state elaborate **5114 url** contenute nel sito di cui 4887 con successo, 22 con errori e 267 ignorati.

Sono stati anche censiti link a 267 host esterni al sito.

Il processo di harvesting ha permesso di archiviare la stragrande maggioranza dei contenuti del sito, anche se in alcuni casi non è andato a buon fine il download di file pdf raggiungibili da link interni al sito.

La dimensione totale del file .warc, prodotto nel processo di harvesting, è di 3.6 GB, suddiviso in quattro file secondo le regole del formato che prevedono una dimensione massima di ogni singolo file di 1 GB.

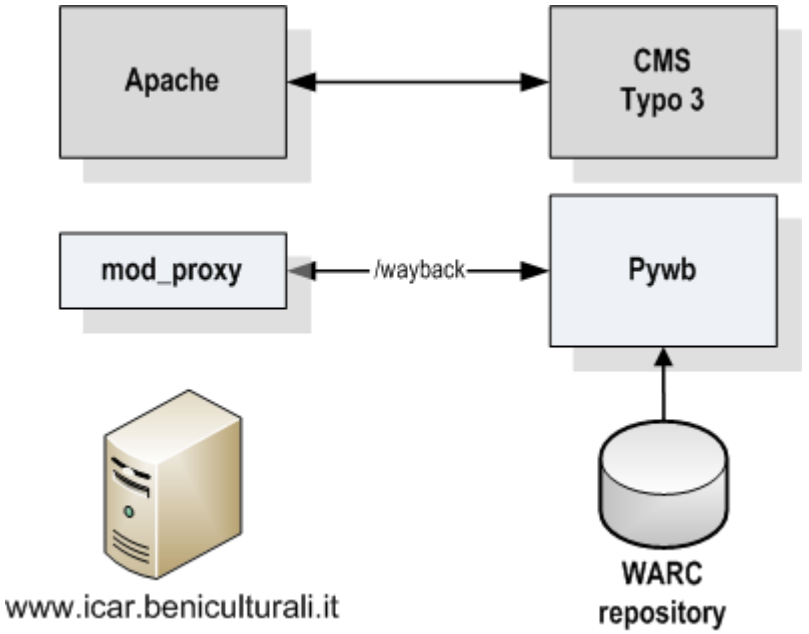
Integrazione nel sito ICAR

La seconda fase del progetto è stata finalizzata a integrare il software Pywb (Python WayBack for web archive replay and live web proxy) nel sito per poter navigare nei file in formato WARC compresso.

Il software Pywb è stato configurato per lavorare in parallelo con il **cms TYPO3**.

E' stato impostato il reindirizzamentod elle url "/wayback "con un modulo proxy .

Le pagine web di Pywb sono state ersonalizzate nei colori dei nuovi loghi e dell'aspetto grafico attuale del sito dell'Istituto.



ICAR
ISTITUTO CENTRALE PER GLI

Demo: Integrazione con sito ICAR

La conservazione del sito ...

www.icar.beniculturali.it/index.php?id=235

Più visitati Come iniziare

PEC: mbac-ic-a[at]mailcert.beniculturali.it e-mail: ic-a[at]beniculturali.it tel: (+39) 06 5196.0286

Istituto Centrale per gli Archivi - ICAR

Cerca...

ISTITUTO | SISTEMI E PORTALI | STANDARD | ATTIVITÀ E PROGETTI | BIBLIOTECA ON-LINE

Home > Attività e progetti > Progetti ICAR > La conservazione del sito web dell'Istituto centrale per gli Archivi (2008-2016)

Il sito Icar nell'Internet Archive

La scelta del software per l'harvesting

La scelta del sistema di visualizzazione e navigazione

La conservazione del sito web dell'Istituto centrale per gli Archivi (2008-2016)

Il settore dei beni culturali ha prodotto e continua a produrre una quantità sempre maggiore di contenuti digitali che hanno bisogno di essere archiviati, conservati e tutelati nel tempo in modo affidabile per consentire che queste risorse possano essere recuperate e riutilizzate in maniera efficace e possano essere utilizzate per le future ricerche storiche.
Le problematiche relative alla **conservazione del patrimonio digitale** non sono state prese in sufficiente considerazione nella stragrande maggioranza delle iniziative di digitalizzazione del patrimonio e in quelle di costruzione di contenuti culturali, per cui di frequente si assiste alla loro scomparsa o impossibilità d'uso dei siti web, con la conseguente perdita della loro valenza culturale e storica e dispendio di risorse umane ed economiche.
Consapevole dell'importanza della **conservazione dei siti web**, quali serbatoi di contenuti culturali e testimonianza storica dell'attività e delle strategie di comunicazione dei soggetti che li hanno promossi, l'Istituto Centrale per gli Archivi, nell'occasione della realizzazione di un nuovo sito web, aggiornato nelle tecnologie e nei contenuti, ha ritenuto necessario affrontare il problema di conservare quello precedente, nella sua interezza e nelle sue varie componenti (pagine web, oggetti digitali, basi di dati), applicando i principi, le metodologie e le tecniche del **web archiving**, così come si è andato definendo nel corso dell'ultimo quindicennio.
Il progetto, ideato e realizzato nel dicembre 2016 da **Costantino Landino**, collaboratore dell'Istituto, si è articolato in due fasi principali: la prima dedicata all'harvesting e alla conservazione del sito web e la seconda alla predisposizione dei servizi di visualizzazione del sito archiviato.

Vai all'home page del precedente sito web dell'ICAR

ICAR
ISTITUTO CENTRALE PER GLI

Demo: Integrazione con sito ICAR

http://www.icar....eniculturali.it/

www.icar.beniculturali.it/wayback/ICARWEB/20161213131014/http://www.icar.beniculturali.it/

Più visitati Come iniziare

ICAR ISTITUTO CENTRALE PER GLI ARCHIVI

Attenzione! Questa è una istantanea del precedente sito ICAR alla data del 13 Dicembre 2016.

Home | Ricerca | Mappa del sito | Crediti | Contatti

- L'Istituto
- Attività
- Circolari e documentazione
- SAN - Newsletter
- SIAS - Sistema informativo degli Archivi di Stato
- Normativa
- Standard e linee guida
- Bibliografia, interventi e materiali didattici
- ICAR. 500 giovani per la cultura. Progetto formativo
- Biblioteca digitale
- Siti
- Osservatorio
- Bandi, gare e contratti
- Trasparenza, valutazione e merito

Home

Finalità
Organismo di studio e ricerca applicata della Direzione generale per gli Archivi, l'Istituto, nell'ambito della sua Attività, é responsabile della gestione, manutenzione e sviluppo dei sistemi informativi archivistici (Sistema Archivistico Nazionale - SAN, aggregatore nazionale di risorse archivistiche; Sistema Archivistico Statale – SAS; Sistema Informativo degli Archivi di Stato - SIAS; Guida Generale degli Archivi di Stato italiani); elabora metodologie in materia di ordinamento e inventariazione di archivi storici, gestione e conservazione degli archivi in formazione, applicazione di nuove tecnologie; sviluppa piani e programmi finalizzati alla descrizione archivistica, alla normalizzazione dei criteri di descrizione, allo sviluppo e all'interoperabilità fra sistemi informativi; cura l'elaborazione di linee guida e standard per l'acquisizione, il trattamento e la gestione delle immagini, l'interoperabilità tra sistemi informativi; promuove l'integrazione e condivisione delle risorse archivistiche informatizzate, la conoscenza e l'applicazione di standard descrittivi e tecnologici, la cooperazione tra istituti archivistici.

Circolari e documentazione
La sezione ospita Circolari di indirizzo emanate dall'ICAR in ragione della specificità delle proprie competenze e Documentazione tecnica con particolare riferimento ai sistemi SAN (Sistema Archivistico Nazionale - SAN), SIAS (Sistema informativo degli Archivi di Stato), SAS (Sistema Archivistico Statale), SIUSA (Sistema Informativo Unificato Soprintendenze Archivistiche), Guida generale (Guida generale degli Archivi di Stato italiani).

Sistema informativo degli Archivi di Stato
L'Istituto ha realizzato, gestisce e sviluppa un sistema informativo di ambito nazionale, denominato SIAS.
Il Sistema offre una descrizione integrata del patrimonio documentario conservato dagli Archivi di Stato attraverso informazioni sugli Archivi di Stato, le loro Sezioni e le sedi di consultazione; sui complessi documentari, sulla loro qualità e consistenza, sulle loro relazioni con i soggetti produttori e con gli inventari esistenti; sui soggetti produttori, istituzioni ed enti, persone e famiglie

News

- 25/11/2016
  30 Novembre Convegno "Sistemi informativi archivistici locali e nazionali: lavori in corso ed esperienze a confronto".
- 25/11/2016
  Invito manifestare interesse per l'affidamento di servizi "Supporto all'utilizzo del software open source Archimista all'interno degli Archivi di Stato e delle Soprintendenze Archivistiche e Bibliografiche".
- 16/11/2016
  Interruzione dei servizi sabato 19 e domenica 20 novembre
- 26/04/2016
  Roma, 9 maggio 2016 | Tracce di memoria

Web Archiving: criticità

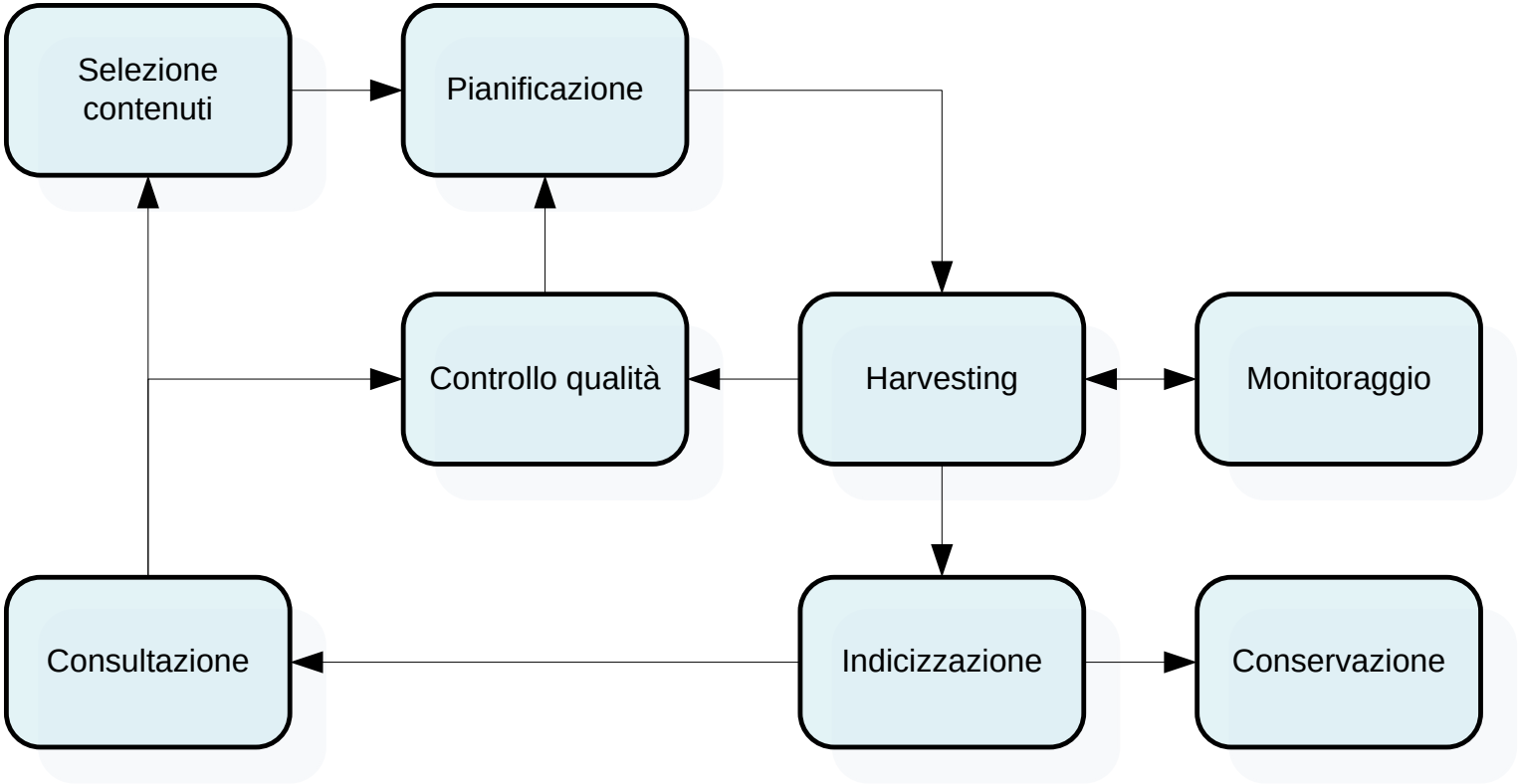
Alcuni problemi rimangono aperti e saranno oggetto di approfondimenti e studi ulteriori.

In particolare:

- La selezione dei contenuti da sottoporre ad harvesting
- Studio ed implementazione dei metadati di accesso e ricerca
- Il controllo qualità
  - l'analisi completa dei link harvestati
  - la correzione dei broken link segnalati dall'errore "404 page not found”
  - la gestione dei link esterni al dominio originale
- L’implementazione delle soluzioni di harvesting
- La complessità di una eventuale indicizzazione

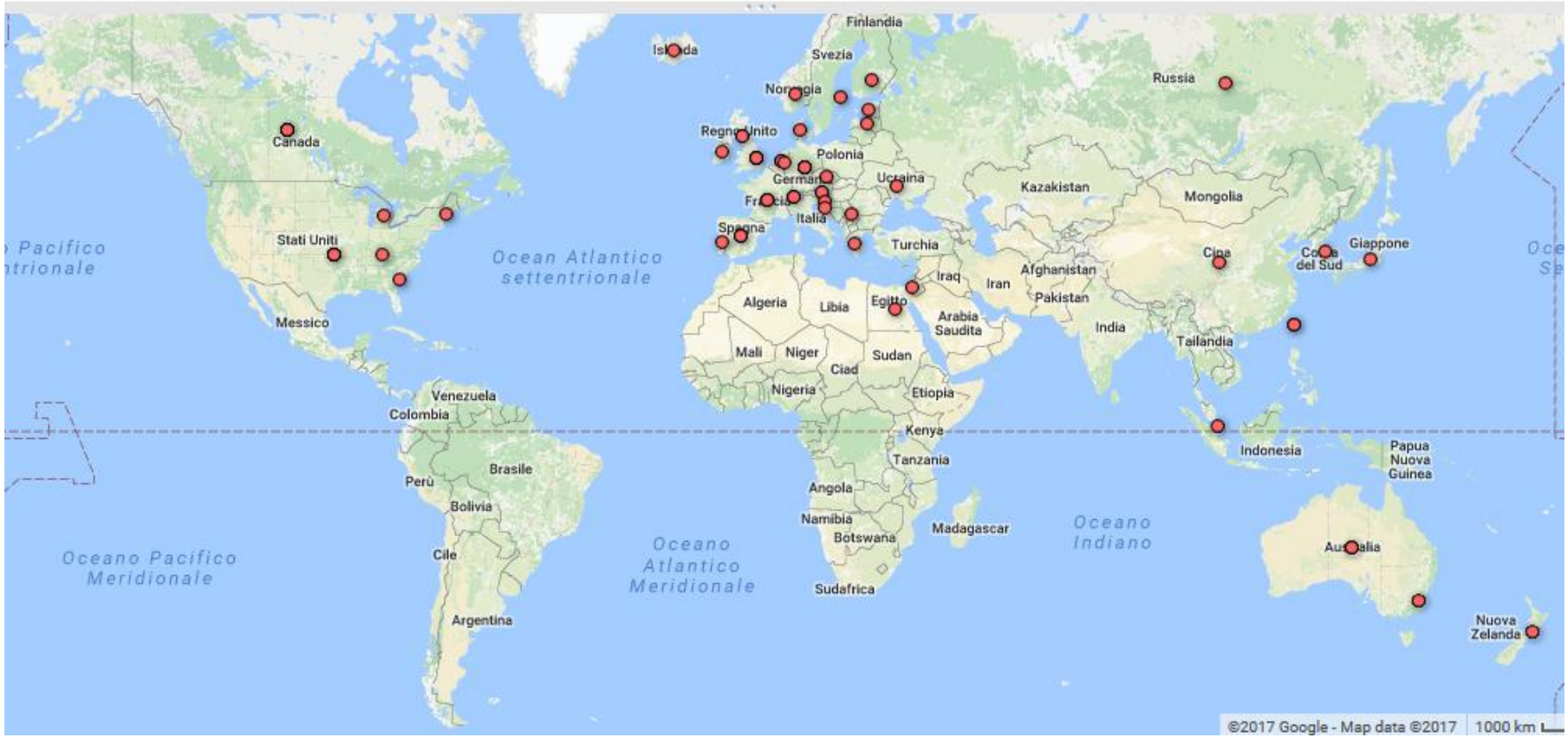
Web Archiving: processo rivisto

A conclusione del lavoro, il processo va rivisto per tenere conto di alcune criticità emerse: la selezione dei contenuti, il controllo qualità e l'indicizzazione.



ICAR
ISTITUTO CENTRALE PER GLI ARCHIVI

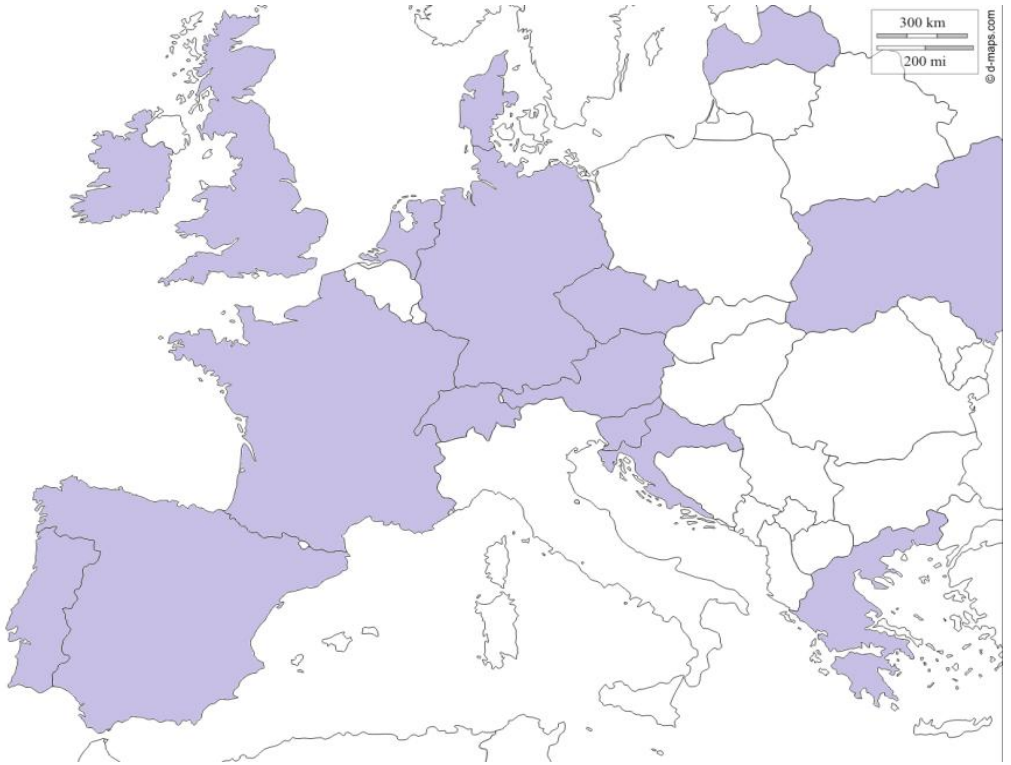
Le iniziative di WEB Archiving nel Mondo



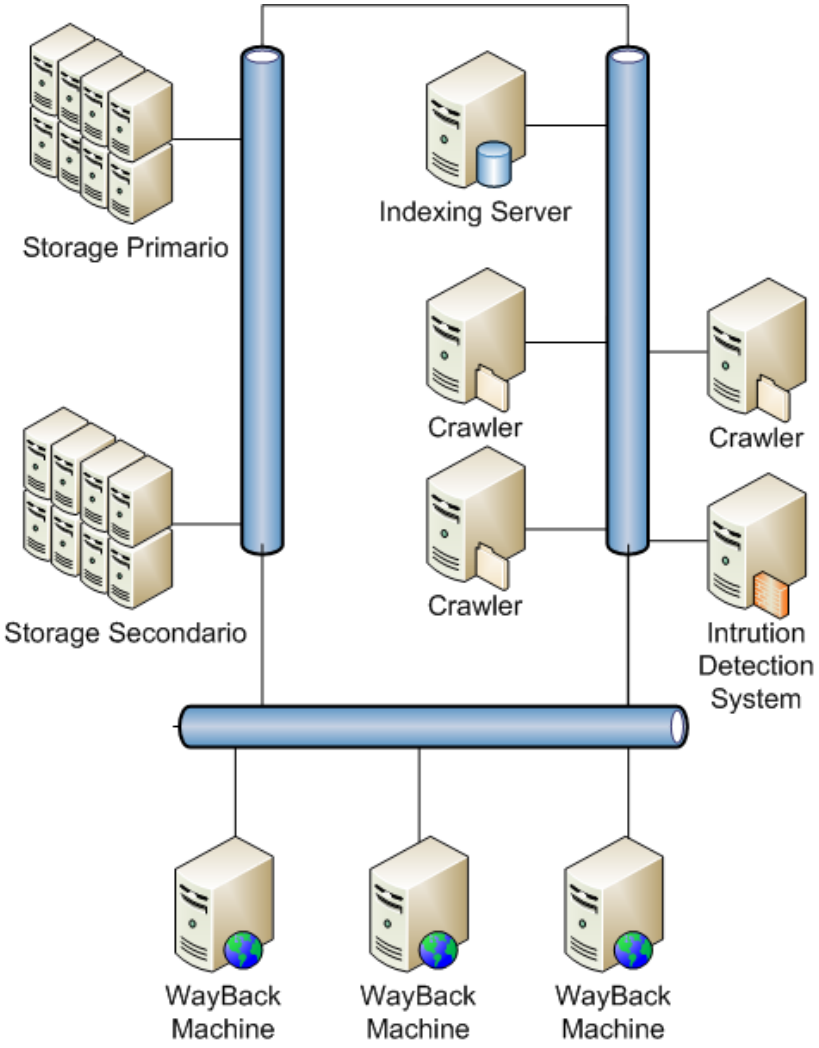
Le iniziative di WEB Archiving in Europa

La maggior parte degli **Archivi** e delle **Biblioteche Nazionali Europee** ha intrapreso progetti di **web archiving** delle **risorse più significative** del proprio paese.

Sono stati creati portali nazionali dedicati in tutti i paesi europei.



Web Archiving: architettura complessa



Web Archiving: sfida

Conservare le nostre memorie

Grazie!

Costantino Landino

costantino.landino@beniculturali.it
costantino.landino@gmail.com

ISTITUTO CENTRALE PER GLI ARCHIVI
www.icar.beniculturali.it
Direttore: Stefano Vitali (ic-a.direttore@beniculturali.it)