THE SYNERGY

REFERENCE MODEL OF

DATA PROVISION AND AGGREGATION

Draft v1.5

July 2016

Current Contributors: Martin Doerr[1], Achille Felicetti, Gerald de Jong[2], Konstantina Konsolaki[1], Barry Norton[3], Dominic Oldman[3], Maria Theodoridou[1], Thomas Wikman[4],

[1]Institute of Computer Science, FORTH
{martin, konsolak, maria@ics.forth.gr}

[2]Delving B.V.
{gerald, thomas@delving.eu}

[3]British Museum
{mailto:Bnorton, DOLDMAN@britishmuseum.org}

[4]The National Archives (Riksarkivet)
{thomas.wikman@riksarkivet.se}

TABLE OF CONTENTS

List of Figures

List of Tables

1 Introduction

This document defines a new reference model for a better practice of data provisioning and aggregation processes, primarily in the cultural heritage sector, but also for e-science. Such processes have become a reality in various, largely disharmonious, forms and in more or less systematic ways in numerous publicly funded projects. The current document defines a consistent set of business processes, user roles, generic software components and open interfaces that will form a harmonious whole. It is based on experience and evaluation of national and international information integration projects. The model is an initiative of the CIDOC CRM Special Interest Group, a Working Group of CIDOC-ICOM, the International Committee for Documentation part of the International Council of Museums. This document is a first draft compiled by a panel of Group members to initiate discussion and further elaboration by the Group and all interested experts and stakeholders in the area. This draft is a skeleton and is in no part complete or elaborated to the intended level of detail and does not represent any authoritative decision or approval of content. The contributors hope that this document is specific and elaborate enough that the reader can appreciate the scope, utility, form and level of specificity of the intended model so that they collaborate in an informed manner. Any and all interested experts are invited to participate and contribute.

The rationale behind this model is the following: Managing heterogeneous cultural heritage data is a complex challenge. Member institutions like galleries, libraries, archives and museums curate different types of collections that, even between similar types of institutions, are documented in different ways using different languages; influenced by different disciplines, objectives and geography, and are encoded using different metadata schemas. However, handling these metadata as a unified whole is vital for progressing new fields of humanities research and discovery, providing more knowledgeable information retrieval and (meta) data exchange, and advancing the field of digital humanities in its various aspects.

The ability to provide users with a uniform interface to access, relate, and combine data while at the same time preserving all the meaning and perspective of the individual data providers, might at first seem like an impossible task. Indeed, the exponential growth of the Web and the extended use of database management systems has brought to the fore the need for the seamless interconnection of large numbers of diverse information sources. In order to provide such uniform access to such heterogeneous and autonomous data sources, complex query and integration mechanisms need to be designed and implemented.

Data aggregation and integration has the potential to create rich resources useful for a range of different purposes, from research and data modeling to education and engagement. There are now significant numbers of projects that aggregate data with these purposes in mind. However, aggregators face two problems.

Firstly, the process of transferring data from source institutions to a central repository can result in a form of data representation stripped of essential information and institutional perspectives. This occurs when mandating target models into which all data sources must fit, regardless of their range, individuality and richness. The generalizations used in these models, designed to facilitate data integration, are too abstract to support the meaningful connections that undoubtedly exist in the data and thus significantly reduce the value of such aggregation initiatives.

The second problem, addressed by this document, relates to the lack of sustainability in the mechanisms and processes through which data is mapped, and the weakness of the partnership between data providers and aggregators inherent in these flawed processes. The mechanisms used for transferring data do not include the full set of necessary processes and tools to create a consistent and good quality outcome and furthermore cannot practically respond to changes in schema and systems on either side of the data provisioning relationship. In order for systems to be sustainable a broader approach is needed that incorporates the experience and knowledge of provider institutions into the infrastructure, in an accessible, beneficial and cost effective manner.

Therefore this document describes a new data provisioning model, the "**Synergy Reference Model**" (specifically the provision of data between providers and aggregators) including associated data mapping components. The intention is to address the design flaws in current models and crucially incorporate, through additional processes and components, the required knowledge and input from providers to create good quality, sustainable aggregations. The funding allocated to humanities aggregation projects over the last two decades has not generated the benefits and progress enjoyed in other sectors who have taken better advantage of digital innovation by using solid and inclusive infrastructures. Unless the value of these infrastructures is clearly demonstrated in the cultural heritage sector, resources for building and developing them will become even more scarce. Unfortunately, numerous systems initiated by projects in both Europe and the United States have failed to understand and identify the relationships and activities necessary to operate collaborative aggregation systems properly and instead have relied on one-sided and centralized approaches using top down modeling and technology led solutions.

The impact of these unresolved issues has been highly detrimental to the landscape of digital humanities, which has thus failed to evolve sufficiently from its fragmented beginnings. Consequently progress has not been made on the essential infrastructure necessary to produce meaningful innovation and fulfill the domains ambitions. Current projects still focus on short term functionality rather than tackle the key issues of sustainability and longevity. Together with an equally debilitating lack of an expressive and real world cultural heritage reference model for data, the humanities has not come close to tackling the challenge of computational reasoning or sophisticated modeling techniques across their vast heterogeneous resources. These resources still remain unexplored and effectively siloed, even when included in aggregation repositories. Indeed such reasoning and modeling activities hold the key to the serious advancement of humanities research; to a degree that would be of serious interest to funding bodies, who are instead becoming increasingly impatient and dissatisfied.

Aggregation systems cannot be viewed as substantially centralized undertakings because aggregators divorced from their providers cannot represent data properly and therefore cannot provide data integration with any real value to any audiences, or, crucially, to the providers themselves. As a result providers are unwilling to commit resources beyond their allocated project funding. If aggregation systems are to have any value they must distribute roles and responsibilities and include the experts who understand the data and know how it should be represented. On the other hand, providers should have the business interest to keep their data consistent and up to date. Data providers curate and understand their information. Aggregators must restrict their involvement to providing homogeneous access to integrated data in such a way that it retains its proper context and its original meaning. Only then can the aggregator provide meaningful services to providers and users. The aggregator provides the information processes (including data improvement and co-referencing) that individual organizations are unable to resource independently, and generate quality services that can be utilized (in return for their data) by the provider. The provider also benefits, if the aggregation is done well and conveys the full meaning of the data, from the ability to use these integrated digital resources to support their own digital strategies. Aggregations based on meaning and context support all organizations small and large and increase the value and relevance of cultural heritage resources for a range of different purposes.

The process of mapping needs support from carefully designed tools and a collaborative knowledge base or “mapping memory” designed to support all organizations with differing levels of resourcing. Together with the CIDOC CRM ontology a new provisioning model is seen as a major step forward in the ability to directly enrich data through collaborative data harmonization, and using the power of multiple data sources to correct, inform and provide greater insight. The challenge

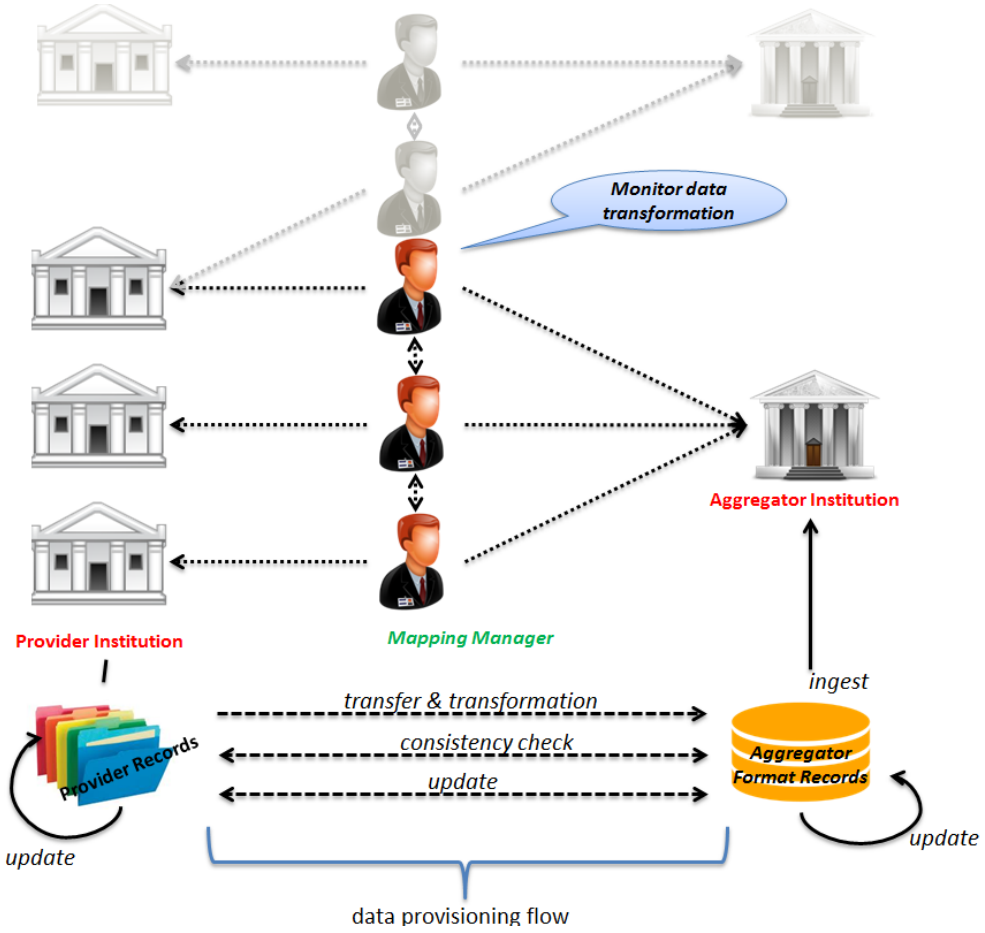

Figure 1: The data provisioning process

is to define a modular architecture that can be developed and optimized by different developers with minimal inter-dependencies and without hindering integrated UI development for the different user roles involved. The first part of this model is a form of requirements specification, which breaks down in the usual way into a definition of the associated user roles, the primary types of data the system aims to handle, and the complete definition of the processes users of the system carry out to manage the data.

The Synergy Reference Model will be described here in terms of a formal process model. The process model requires the definition of the individual roles, data objects and processes necessary for designing a controlled and managed mapping system. Figure 1 presents a high level view of the data provisioning process. Different Provider Institutions may keep a relationship with the same Aggregator Institution, but also one Provider Institution may keep a relationship with more than one Aggregator Institution. The Provider Institution and the Aggregator Institution agree on the data provisioning and related activities. A “Mapping Manager” is nominated by both parties to oversee the actual data transfer process and forms the third primary role in the Synergy Model. The Provider Institutions own the Provider Records which are transformed to the Aggregator Record Format and are transferred

to the Aggregator Institution. The data provisioning process is regarded as an open-ended and on-going task. Throughout the transformation and transfer processes, consistency checks and updates are necessary between all partners and will be supported by the model. The model foresees a series of distinct update processes at all partner sites which trigger each new data transfer. For the sake of simplicity, this model only describes the N to 1 relationship between the Provider and the Aggregator Institution.

The details of the Synergy Model will be presented in this document. As is normal in describing processes the descriptions may contain some redundancy. For example, it is common for different roles to be performed by the same people. The processes presented in this document should not be viewed as complete and static but rather are designed to facilitate the growth of other collaborative processes between provider organizations and aggregators. The ability to refer to the same set of stable processes increases the opportunities for organizations to pool and share their resources whether aimed at improving the aggregation service or simply at providing a platform for collaboration outside or connected to the offered aggregation services.

The structure of this document is as follows:

In Data Provisioning, we present an overview of the data provisioning process to provide a context for the next sections. In the following sections we show analytical model views[1] needed for a detailed understanding of the reference model. In section User Roles we give a detailed description of all the involved roles while in , we present the data objects that assist the data provisioning process. Data Provisioning Flow Network, presents the flow of the data objects from the Provider Institution to the Aggregator's Institution. It also provides a high-level overview of the IT objects that either replace manual tasks or assist the user. Analytical Representation of the Data Provisioning Process, presents the detailed analysis of the data provisioning processes. Finally, Services and Software components presents in detail the IT objects that assist the mapping process. All sections are under construction, and only indications of the intended final content.

[1] The modeling of the data mapping components and also of the processes needed for the completion of the mapping is made using Adonis-Business Process Management [1]. Adonis is a freeware tool, useful for the design and documentation of processes.

2 Data Provisioning

The Synergy Reference Model aims at identifying, supporting or managing the processes needed to be executed or maintained when a provider and aggregator agree (see):

1. to ***transfer*** *data* from the provider to the aggregator,
2. to ***transform*** their format to the (homogeneous) format of the aggregator,
3. to ***curate*** the semantic consistency of source and target data and the global referential integrity and
4. to ***maintain*** the transferred data *up-to-date* with whatever relevant changes occur in the source and target systems and the employed terminologies.

In the following we present the requirements and assumptions taken into account during the Synergy Model design.

2.1 Requirements and assumptions

The Synergy Reference Model aims to support the management of data between source and target models and the delivery of transformed data at defined times, including updates. This includes a mapping definition, i.e., specification of the parameters for the data transformation process, such that complete sets of data records can automatically be transformed. A provisioning model includes:

1. The ***transfer of data*** (and iterative corrections) until a first consistent state is achieved. This includes transformation of sets of data records submitted to the aggregator, the necessary exception processing of irregular input data between provider and aggregator, ingestion of the transformed records into target system and initial referential integrity processing possibly on both sides.

Note: Referential integrity processing at the aggregator side, is out of the context of a particular data submission, is a necessary process but is not part of the model.

2. The ***change detection and update*** processing to provide updated information and ensure semantic consistency in situations where changes have been made in the source or target records, in the provider or aggregator terminologies, in the source or target schemata, in the target instance generation policies and changes in the interpretation of source and target schema recorded in the mapping definition.

It is assumed that the provider has mechanisms that would identify modified records and thus ask for the mapping of these records only. If this is not the case, either the whole mapping is executed, overwriting previous mappings, or smaller units of change are identified and updated. Additionally the target system may also require

some way to clearly identify a modified record. Some aggregation systems may wish to store versions of data for research purposes but in this case the canonical records should always be clear and differentiated. As long as these changes can be identified and accessed then it is a matter for the aggregator to determine their own versioning system.

Only if these processes are sustained can an aggregator provide valid and consistently integrated data over the long term, and thereby deliver the full added value of an aggregation service that would make it attractive for providers and users alike. This sustainability is key in providing benefits for providers and the range of professionals, experts and enthusiasts that would ultimately justify its existence. This report is aware that none of the hundreds of mapping tools and frameworks created in numerous projects has ever systematically addressed this comprehensive scenario.

It is not until data can be analyzed and visualized in an appropriate format and environment that mapping decisions and issues can be made. The system should allow the exploration of data and give some indication of inconsistencies that might exist. The visualization process also provides the ability for both source and target models to be compared. The system may also allow ‘test’ transformations and provide some of the functionality, described below, to be applied to individual cases to provide some understanding of the requirements necessary to complete a full mapping and transformation.

Only if these processes are sustained can an aggregator provide valid and consistently integrated data over the long term, and thereby deliver the full added value of an aggregation service that would make it attractive for providers and users alike. This sustainability is a key in providing benefits for providers and the range of professionals, experts and enthusiasts that would ultimately justify its existence. This report is aware that none of the hundreds of mapping tools and frameworks created in numerous projects has ever systematically addressed this comprehensive scenario.

It is not until data can be analyzed and visualized in an appropriate format and environment that mapping decisions and issues can be made. The system should allow the exploration of data and give some indication of inconsistencies that might exist. The visualization process also provides the ability for both source and target models to be compared. The system may also allow ‘test’ transformations and provide some of the functionality, described below, to be applied to individual cases to provide some understanding of the requirements necessary to complete a full mapping and transformation.

3 User Roles

The following subsections describe the key management roles that oversee the process and provide the necessary resources. We distinguish primary and secondary roles associated with the data provisioning process. As depicted in Figure 2, the provider and aggregator organizations consist of performers, the *provider* and the *aggregator employees*, that may be one person or a team. Since the model does not prescribe if the mapping process is managed on account of the provider, the aggregator, or a joint activity, we have named this role as "*provider or aggregator employee"*, that may be one person or a team that is employed by the provider institution or the aggregator institution or both. Each performer holds different roles during the data provisioning process and as depicted in Figure 2 one performer may have in common one role with another performer. It is obvious that a person can play more than one roles in such a reference model.

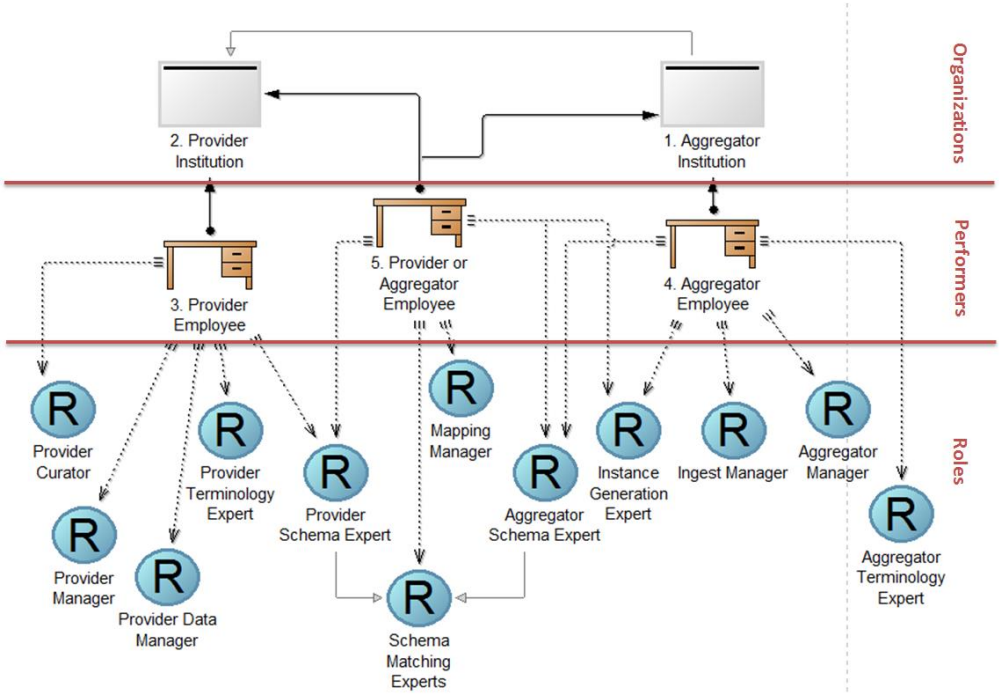

Figure 2: Working Environment - User Roles

3.1 Primary User Roles

As primary user roles we regard the managerially responsible members from the Provider and the Aggregator Institution that agree to perform the data provisioning

from the providers' local information systems to the aggregator's integrated access system. These roles are seen in a logical and not a personal way. The model does not describe or exclude that the institution maintaining an aggregation service may also maintain a provider service. In that case, the provider role applies to the respective functions of such an institution. It is not part of the model where the actual data physically reside and how they are replicated or preserved. The model currently does not describe or exclude that an aggregator may hand over data to another aggregator. In that case a chain of interactions should be installed which ensures an information flow between primary providers and primary and secondary aggregators functionally equivalent to the one proposed by this model.

We define the following primary user roles:

Provider Institution

The ***Provider Institution*** maintains one or more Local Information Systems. Following CIDOC CRM v5.0.4 [2]:

"*These are either collection management systems or content management systems that constitute institutional memories and are maintained by an institution. They are used for primary data entry, i.e. a relevant part of the information, be it data or metadata, is primary information in digital form that fulfils institutional needs.*"

In practice these are systems owned by individual museums, archives, libraries, sites and monument records, academic institutes, private research societies etc., represented by their curators, technologists, documentation personnel and researchers. In other words, "Provider" in the sense of this model is the role of an institution as the authority for the correctness of knowledge represented in the data. Provider Institution systems are also called **source systems** in this text when talking about data transformation and submission. The implementation of provider systems is not part of this model, only certain communication capabilities.

Aggregator Institution

The ***Aggregator Institution*** maintains an Integrated Access System and in this model may also be called simply, 'target systems'. Following CIDOC CRM v5.0.4 [2] "*These provide a homogeneous access layer to multiple local systems".* The origin of the information it manages are the Provider Institutions it maintains a business relation with. It may not alter provided content except for co-reference resolution information, changes of identifier and value representations or schema migrations. It may remove provided content or ask to providers for updates in order to maintain data quality. In case the aggregator institution wishes to add own new content of any form and provenance, the model will treat this part of information as another source system and will regard the respective activities as Provider activities. In other words,

“Aggregator” in the sense of this model is the role of an institution of integrating and mediating data without changing meaning.

Aggregator Institution maintains a form of business agreement with providers to send data from local systems to the aggregators’ system, consisting primarily of metadata. The scenario that aggregators may harvest provider information without any formal terms of reference and understanding with the provider, such as the well-known search engine services, is not part of this model. The model will still be of use for such scenarios in a trivial way, but this scenario involves activities that are not the focus of this document and implies aggregators that have no direct knowledge about the meaning of the data they aggregate. This is insufficient for services that seek to harness the rich nature and embedded knowledge of cultural heritage organizations.

Mapping Manager

The ***Mapping Manager*** is the actor responsible for the maintenance of the data transformation process from the provider format to the aggregator format. This role may split into a semantic and a technical part, and may be regarded as an aggregator task, a provider task or a user consortium task. The mapping technology, this model aims at, should support scalable management of the data transformation process by the aggregator. Mapping Managers must schedule and negotiate the terms under which data transformation occurs at both ends of the data provisioning system realizing that this may differ between providers.

3.2 Secondary User Roles

In this section we describe the experts whose knowledge or services contribute to the implementation and realization of the data provisioning process. They hold the secondary roles in the Synergy Model and they are employees of either the Provider or the Aggregator Institution.

Provider Curator

The Provider Curator is the employee responsible for curating the content of the source systems. They have the understanding of how their data reflects the real world, or know those who do, and have the means to check the veracity of the representation.

Provider Manager

The Provider Manager is the employee responsible for negotiating and supervising the business agreement established with the Aggregator's Integrated Access System.

Provider Data Manager

The Provider Data Manager is the employee responsible for managing the relevant IT systems of the provider and for handling the data assets, in contrast to those responsible for creating content. In particular, they are responsible for sending data to the aggregator. The Provider Data Manager may formally be employed by the Provider Institution or the Provider Institution has outsourced this service. For the purpose of this model, it is not necessary to differentiate these situations.

Provider Schema Expert

The curator(s), researcher(s) and/or data manager(s) of the Provider Institution who are responsible for the data creation in their local systems, i.e., the people who know how, following local use and practice, the fields, tables or elements in the schema correspond to the reality described by them. These meanings may have become skewed or misinterpreted over time (the so-called semantic drift). This can be caused by a lack of precision or differentiation in the underlying data models, through misrepresentation in overlaying software and user interfaces or through changing practice. As such these meanings have to be researched and understood before mapping can take place. This is the domain of an increasingly undervalued group of people responsible for the quality and semantic correctness of their data. This group's significance and value would be highlighted by a real, rather than technically artificial, representation of their work. While quality levels will vary from institution to institution, semantic data *harmonization* can contribute to improving the quality of information across all organizations and indeed the whole sector.

Provider Terminology Expert

The Provider Terminology Expert is the curator, maintainer, or other expert on one or more of the terminologies that the provider uses as a reference in the local system. If the terminology is provided by a third party, such as the Getty Research Institute, there may exist independent external experts conversant with this terminology. If the terminology is local in origin, or even uncontrolled, it is typically the curators or other local data managers (and documentation staff) who, alone, know the meaning of the local terms.

Aggregator Schema Expert

The Aggregator Schema Expert is the expert on the semantics of the schema employed by the aggregator ("integration model"). Some large scale aggregators use a more widely known standard schema, but there is also a growing trend in the linked data world towards lightweight portal or 'indexing' aggregation projects that implement narrow custom models. This document refers to aggregation using data modeled according to intelligent frameworks which incorporate the possibility of a wider context and are based on cross disciplinary expert knowledge and description. Typically but not exclusively, this document refers to the CIDOC CRM and extensions of it. The CRM provides a richer but smaller entity model compared to the source models that are mapped to it. An identified issue is that aggregators, despite having expertise in the use of the target schema, have significant gaps in their knowledge about established curatorial practices. This situation can lead to a mismatch of semantics between the provider and the aggregator and therefore the data provision process assumes a greater level of contact with provider representatives than is currently the case. This relationship has a direct effect on increasing the quality of the aggregation.

Aggregator Terminology Expert

The Aggregator Terminology Expert is the curator, maintainer or other expert on one or more of the terminologies that the aggregator uses as reference in the Integrated Access System. Aggregators normally want to avoid engagement with terminology maintenance. They often use more generalized provider independent terminologies rather than (more rarely) take over provider term lists. However, it should be noted that the richness of an aggregation may include the ability to understand why different terminology has been applied in different organizations and that this is also a means by which things can be re-contextualized, to a certain extent, to the real and historical world. Established local terminology should never be replaced by centralized and generalized terminologies, but be correlated to the latter.

Ingest Manager

The Ingest manager is responsible for receiving data from the provider and ingesting them into the target system.

Aggregator Manager

The Aggregator Manager is the employee responsible for negotiating and supervising the business agreement established with the Provider's Institution.

Instance Generation Expert

The expert of the aggregator, normally an IT specialist, who is responsible for maintaining the referential integrity of the (meta)data in the Integrated Access System and who knows how to generate from provider data valid URIs for the Integrated Access System.

Schema Matching Experts

Provider schema experts and an aggregator schema expert collaborate in order to define a schema matching, which is documented in a schema matching definition.

4 Data Objects

We define three categories of information objects that take part in the data provisioning process :

1. the ***content data and metadata objects***
2. the ***control objects***
3. the ***schema and logic objects***

The data objects are illustrated in Figure 3 and are described in detail in the following sections.

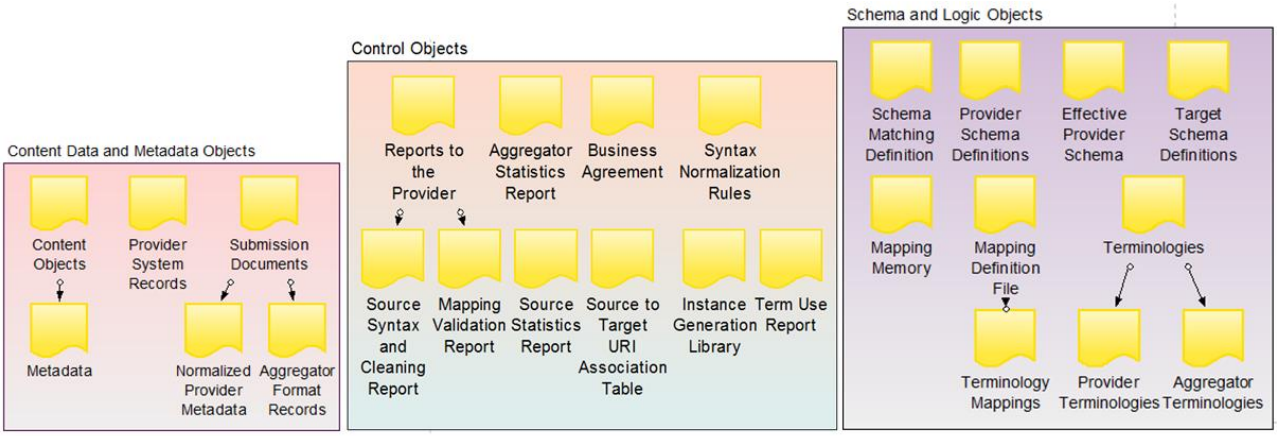

Figure 3: Data Objects

4.1 Content Data and Metadata Objects

Content data and metadata objects consist of all the raw data and metadata. In detail they include:

Content Objects

Individual files or information units with an internal structure that is not described in terms of schema elements of the source or target systems. These are things like images or text documents that are searched with content retrieval indices such as keyword searches, rather than by associative queries. As objects they are typically described by metadata records which are searched by associative queries. What is important in this context is that a content object is not identified by the actual structure it contains but by the way it is treated in the information system (stored either as “blobs” in the database or using references to a file system). Many aggregators do not collect content objects but only link to those resources in the provider system. If these objects are to be collected then they should be referenced

in the mapping and transferred to the target system with the appropriate URL specified in the mapping.

Metadata

Metadata are information units with an internal structure that is described in terms of schema elements of the source or target systems. In our context, these are often data records describing a content object (therefore the term “metadata”). However, bad analogy has also brought the term into use for data describing physical objects. Therefore we define “metadata” here in the same way it is treated in the information system, and not as “data about data”. The metadata records are the common subject of submission to aggregators and therefore of transformation from the source to the target schema.

Provider System Records

The Provider System records are records of the Local Information System. In case they contain fields with local, informal or uncontrolled internal syntax, such as frequently occurring for commented dates, dimensions, or uncertain information, we speak about **Metadata**.

Submission Documents

The Submission Documents are well-formed documents that are produced either by the syntax normalization process, or by the transformation process.

Normalized Provider Metadata

Normalized Provider Metadata is the result of formalizing (“cleaning”) Metadata by extending the provider schema. The result is completely structured data in the effective provider schema, in which each structural element must have a clearly described meaning, i.e., all local methods to subdivide a name or string into meaningful subsections should be expressed by explicit, unambiguous tagging, preferably in XML.

Aggregator Format Records

The Aggregator Format Records are the records in the form to be ingested into the target system.

4.2 Schema and logic objects

Schema and logic objects consist of the schemata, mappings and terminologies of both the source and target systems. In detail they include:

Schema Matching Definition

The Schema Matching Definition contains the mappings of the source schema elements to the target schema paths. This definition must be human and machine readable and is the ultimate communication means on the semantic correctness of the mapping.

Provider Schema Definitions

These include data dictionaries, XML schemata, RDFS/OWL files etc. describing the data structures that are managed and can be searched by associative queries in the source system.

Effective Provider Schema

The Effective Provider Schema is the extended provider schema definition that comes up, in case local syntax rules exist.

Target Schema Definitions

These include data dictionaries, XML schemata, RDFS/OWL files etc. describing the data structures that are managed and can be searched by associative queries in the target system.

Mapping Memory

A collection of mapping histories of analogous cases collected from the user community.

Mapping Definition

The Mapping Definition comes up by the addition of the instance generation policies to the Schema Matching Definition.

Terminologies

We regard as Terminologies controlled vocabularies of terms that appear as individual data values in the source or target systems and represent *categorical concepts* or “universal”. We do **not** regard reference information about places (gazetteers) and people as terminologies in this document. Matching people and places are regarded as cases of “co-reference resolution” in this document. The term “vocabulary” is **not** used for metadata schemata in this document. Terminologies may be flat lists of words or be described and organized in more elaborate structures as so-called “thesauri” or “knowledge organization systems”, the most popular format now being SKOS. This document distinguishes these structures from an

"ontology", even if the terminology may qualify as such, as long as its use in this context is to provide *data values* and not data structure.

Provider Terminologies

The terminologies used by the provider as a reference in the local system.

Aggregator Terminologies

The terminologies used by the aggregator as a reference in the Integrated Access System.

Terminology Mappings

Terminology Mappings are expressions of exact or inexact (broader/narrower) equivalence between terms from different vocabularies.

In this context, we are primarily interested in the mapping of terms that appear directly or indirectly in mapping conditions of a Schema Matching Definition. In such a mapping condition, a term in the source record is equal to or unequal to a constant, or a narrower term of a constant. This may be expressed in terms of source or target terminology (note that we regard as terminology **only** categorical concepts, not names of particular things, events, places or people).

For instance, take a source schema with a table "Object" and field "object type", to be mapped to the CIDOC CRM. The source schema does not distinguish material from immaterial objects. The target schema we map to however makes the distinction. Then, a source field with value "object type = Vase" may indicate a "Physical Object", and "object type = Image" an "Information Object". As a consequence, the value in the field "object type" determines two alternative interpretations of the table "Object". This is a conditional mapping, in which the mapping of a source schema element depends on the value in some other element of the source record. In general, only categorical terms should affect the schema matching logic.

4.3 Control Objects

Control data objects are the reports and documents that support the data provisioning process and are the products of its different sub-processes. In detail they include:

Business Agreement

A Business Agreement sets the expectations between the provider and the aggregator and helps define the relationship between the two parties. It is the cornerstone of how the provider sets and maintains commitments to the aggregator.

Reports to the Provider

They include reports useful to the provider in order to monitor the result of the various tasks and to announce all individual inconsistencies in the processed source data which the provider may or should correct.

Syntax Normalization Rules

The Syntax Normalization Rules report contains the syntax mappings that came up from the effective schema.

Source Syntax Report and Cleaning Report

The Source Syntax Report and Cleaning Report is the output report of the syntax normalizer. It contains inconsistencies and errors that occurred during the syntax normalization process.

Mapping Validation Report

The Mapping Validation Report is the output report of the Metadata Validating Transformer. It contains errors and inconsistencies that occurred during the transformation process.

Source Statistics Report

The Source Statistics Report is the output statistics of the Source Analyzer, used as input to the Source Schema Visualizer. It contains statistic information useful for understanding the provider schema.

Instance Generation Library

The Instance Generation Library includes the generator policies described as a set of rules.

Term Use Report

The Term Use Report contains the list of terms that appear in mapping conditions at the Mapping Definition.

Aggregator Statistics Report

The Aggregator Statistics Report is the output statistics of the Analyzer. It contains statistic information useful for understanding the target schema.

Source to Target URI Association Table

The Source to Target URI Association Table is the output report of the Metadata Validating Transformer. It contains source to target URI associations.

5 Data Provisioning Flow Network

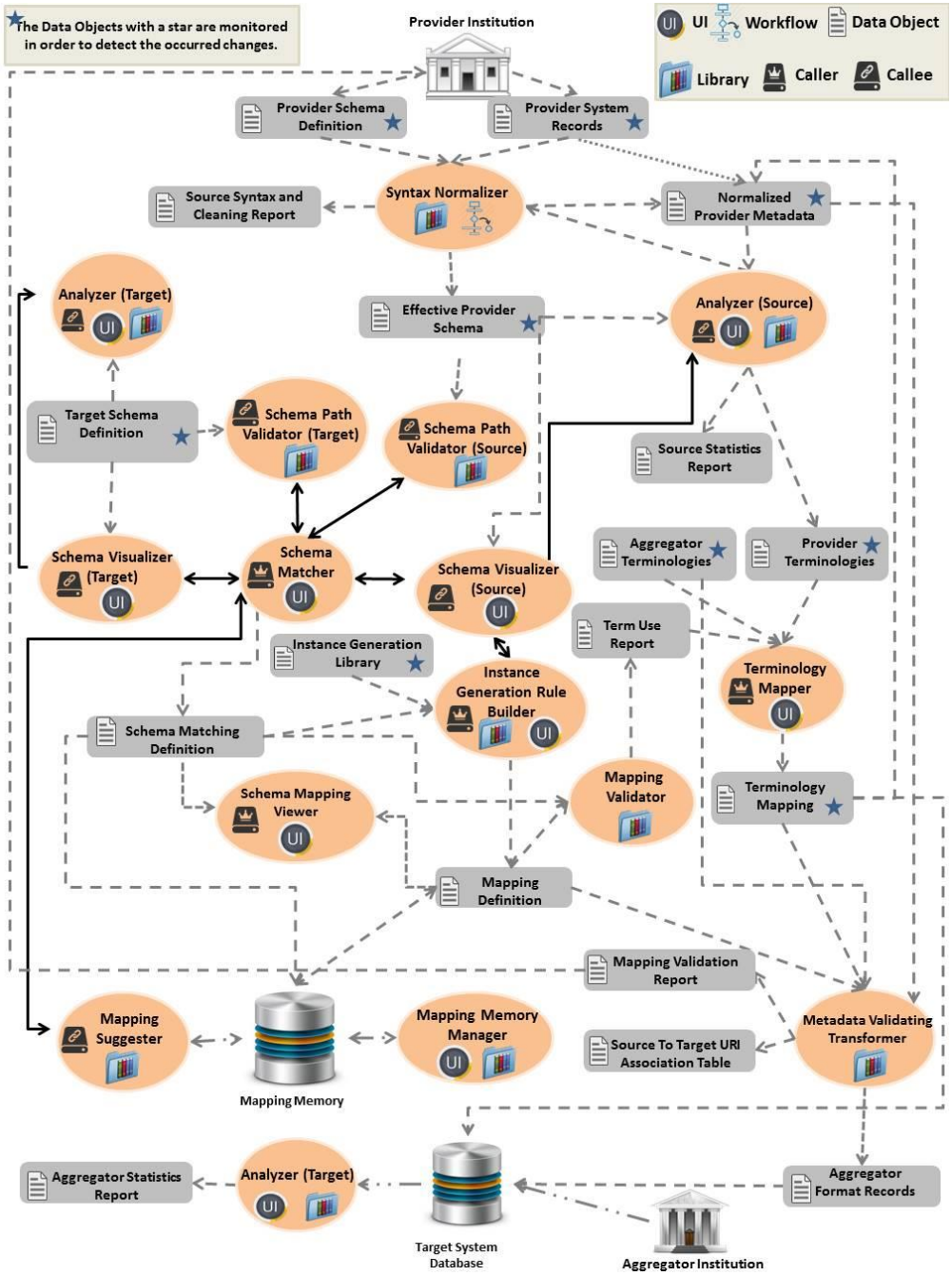

Figure 4: Data Provisioning Flow Network

In this section we present the system objects that assist the data provisioning process and the flow of data between them. The process starts at the Provider Institution and is completed at the Aggregation Institution when all the records are transformed to the aggregator format and are ingested into the target system, as depicted in Figure 4.

Starting from the Provider Institution, the ***Syntax Normalizer*** can be used to normalize the provider's records. It exploits local syntax rules and produces a new provider schema definition, called ***Effective Provider Schema***. Normalization is quite often needed in date fields or in fields that contain concatenated information.

The next step of the provisioning process is the definition of the mappings. Different IT components assist the experts to complete this time-consuming and error-prone task. One of these is the ***Mapping Memory Manager*** (for short ***3M***). 3M is a managing system suitable for handling the mapping files. It offers a variety of actions that help both provider and target schema experts manage their mappings. The first step of the mapping definition is the schema matching. Provider schema experts with target schema experts exploit the ***Schema Matcher*** component, to define a schema matching which is documented in a ***Schema Matching Definition***. The ***Schema Path Validator (Source/Target)*** component assist the experts in selecting the valid paths, with respect to corresponding schemata, preventing them from making mistakes, while the ***Schema Visualizer (Source/Target)*** assist users navigating through all source and target elements. Moreover, the ***Analyzer (Source/Target)*** provides useful statistics for each field. The matching process is also supported by the ***Mapping Suggester***, which makes use of "mapping memories" of similar cases as they are collected and cached from the user community. The ***Mapping Validator*** validates the whole mapping in terms of the source and the target paths.

The next step is the specification of the instance generation rules that define the URI generation policy for each target class. This task is accomplished by the ***Instance Generator Rule Builder***, which complements the Schema Matching Definition with the instance generation policies, producing the ***Mapping Definition***. Both the Schema Matching Definition and the Mapping Definition may be viewed with the ***Schema Mapping Viewer***. The files are stored to the ***Mapping Memory***.

The matching may need to interpret provider and aggregator terminologies in order to resolve data dependent mappings. This may be assisted by a ***Terminology Mapper***.

Finally, when the mapping definition and the terminology mappings are defined, the ***Metadata Validating Transformer*** transforms the records to the aggregator format and ingests them to the Aggregator Institution. Since it is usual for schema documentation to be embedded into the structure of the schema the ***Analyzer (Target)*** should be able to identify this documentation and expose it to the user when browsing aspects of the target.

6 Analytical Representation of the Data Provisioning Process

In this section we will describe in detail the processes that are involved in data provisioning. The modeling of the data provisioning components and processes is designed with Adonis-Business Process Management [1]. The data provisioning process hierarchy is presented in Figure 5.

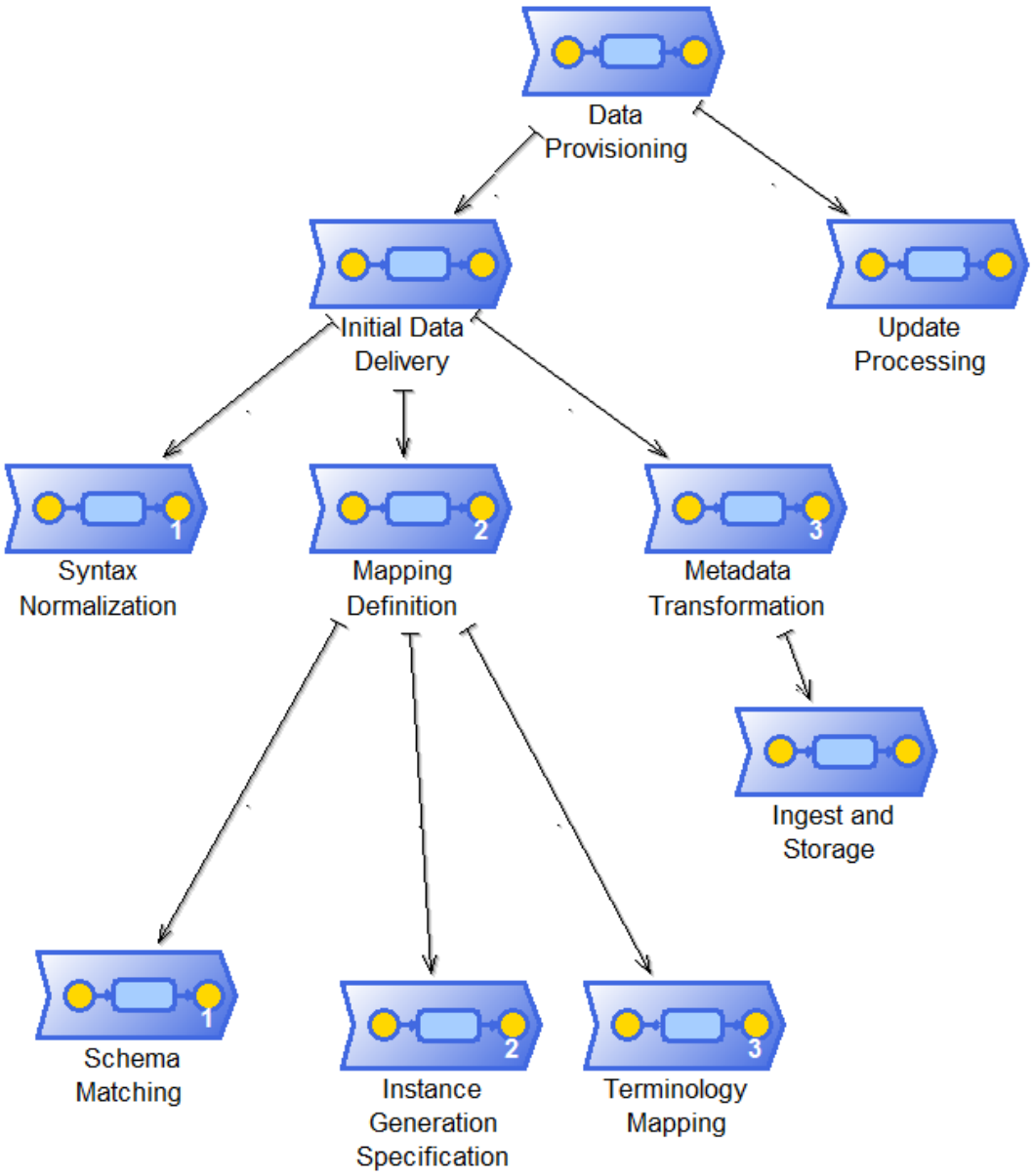

Figure 5: Data Provisioning Process Hierarchy

6.1 Data Provisioning Process

The starting point of the Synergy Model is the Data Provisioning process (Figure 6) which deals with the selection and scheduling of data, including co-reference resolution and updates. A *Mapping Manager* may be responsible for this task that may form part of the agreement with the aggregator that may have been assigned to named representatives on both the provider and the aggregator.

While it is inevitable that the Synergy Model will be unable to deal with all internal provider processes, it should provide general support through the provision of log files and reports. Reports may be based on queries that include changes of provider system records and other information that indicates the need to resubmit data.

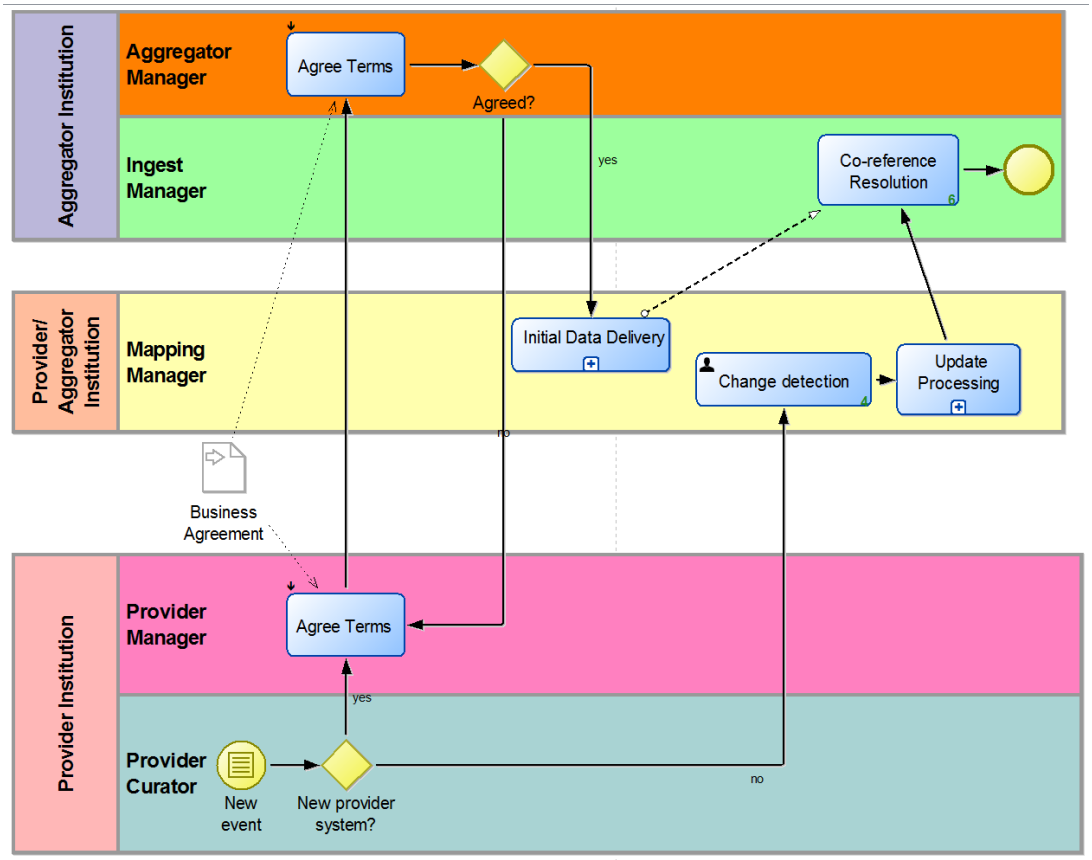

Figure 6: The Data Provisioning process

Queries would also support provider selection criteria in the source systems. On the target system they may be used to reveal semantic needs in the composition of the aggregation and to derive requests for particular or additional materials from providers.

During the data provisioning process, the Ingest Manager may proceed with referential integrity processing (***co-reference resolution***) i.e. resolving multiple identifiers that denote the same real world thing or object of discourse (co-reference resolution). This is a process in its own right and will be the subject of separate documentation. The main goals are to ensure referential integrity, which is the heart

of information integration, and to reduce the number of URIs in use for the same thing. It requires its own dialogue between provider, aggregator and third-party authority managers. Since the aggregator collects more comprehensive knowledge than the providers, it is a natural role of the aggregator. One may regard that the only genuine knowledge of the aggregator, because of the nature of an integrated system, is the co-reference knowledge.

On the other hand, there is a set of characteristic *changes* in the provider – aggregator environment that affect the mapping and may require:

- Re-executing the transformation of records already submitted to the aggregator and updating the transformed records in the target system.
- Resubmission of records from the source system.
- Redefinition of the mapping.

The Mapping Manager must monitor such changes and initiate respective actions. On some occasions, transformation may be affected by a significant change in underlying technology platforms and, while in theory this should not affect the way in which data is mapped to the target schema, such cases may inadvertently prompt changes in the mapping definition and/or require significant re-alignment. This should not be an issue for the system being described.

Table 1 depicts a summary of the tasks presented in **Error! Not a valid bookmark self-reference.**.

Name	Type[2]	Description	Input Document	Role	Organization
New Event	CS	The Data Provisioning process starts only if a new event (a new provider occurs or a change is detected) occurs.		Provider Curator	Provider Institution
New provider system?	EG	Checks if a new provider occurred.		Provider Curator	Provider Institution
Agree Terms	T	The negotiating task between the provider and the aggregator till all commitments are defined and agreed.	Business Agreement	Provider Manager	Provider Institution
				Aggregator Manager	Aggregator Institution
Agreed?	EG	Checks if there is an agreement between the provider and the aggregator.			Aggregator Institution

[2] **CS**: A Conditional Start allows the process to begin or continue only when a business condition or business rule is met.
Sub-p: A Sub-Process is an activity whose internal details have been modeled in a separate process.
T: A Task is an atomic activity within a process flow. It is used when the work in the process cannot be broken down to a finer level of detail.
EG: Exclusive Gateway routes the sequence flow to exactly one of the outgoing branches, when splitting. When merging, it awaits one incoming branch to complete before triggering the outgoing flow.

Initial Data Delivery	Sub-p	The Initial Data Delivery subprocess consists of: 1. The Syntax Normalization subprocess 2. The Mapping Definition subprocess 3. The Metadata Transfer subprocess		Mapping Manager	Provider / Aggregator Institution
Change Detection	T	Monitor the transferred data in order to maintain data up-to-date with whatever relevant changes occur in the source and target systems and the employed terminologies.		Mapping Manager	Provider / Aggregator Institution
Update Processing	Sub-p	Restore ability of data transformation and semantic consistency, which comprises changes in the source or target records, in the provider or aggregator terminologies, in the source or target schemata, in the target URI policy and in the good practice of interpretation of source and target schema in the mapping definition.		Mapping Manager	Provider / Aggregator Institution
Co-reference Resolution	T	Referential integrity processing: resolves multiple identifiers that denote the same real world thing or object of discourse. The main goals are to ensure referential integrity, which is the heart of information integration, and to reduce the number of URIs in use for the same thing. It requires its own dialogue between provider, aggregator and third-party authority managers. Since the aggregator collects more comprehensive knowledge than the providers, it is a natural role of the aggregator. One may regard that the only genuine knowledge of the aggregator is the co-reference knowledge.		Ingest Manager	Aggregator Institution

Table 1: Summary of the Data Provisioning Process

6.1.1 Initial Data Delivery

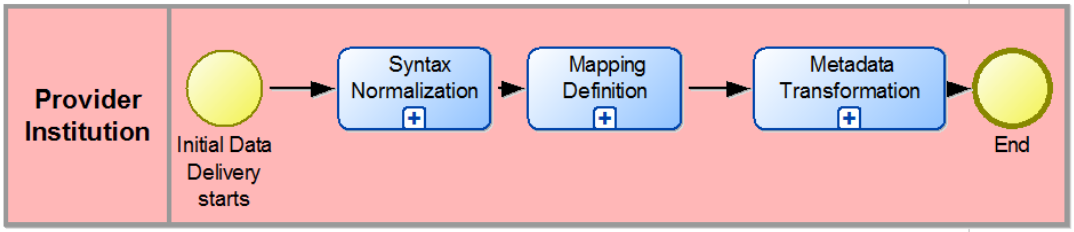

Figure 7: Initial Data Delivery sub-process

Initial Data Delivery breaks down into:

- Syntax Normalization
- Mapping Definition
- Metadata Transformation

Table 2 depicts a summary of the tasks presented in Figure 7.

Name	Type	Description	Organization
Syntax Normalization	Sub-p	Syntax normalization aims to convert all data structures relevant for the transformation in a standard form since data transformation tools can only deal with a limited set of standard data structures.	Provider Institution
Mapping Definition	Sub-p	Mapping definition is the specification of the parameters for the data transformation process, such that complete sets of data records can automatically be transformed, manual exception processing notwithstanding. This includes harmonization between multiple providers.	Provider Institution
Metadata Transformation	Sub-p	The actual transfer of data until a first consistent state is achieved. This includes transformation of sets of data records submitted to the aggregator, the necessary exception processing of irregular input data between provider and aggregator, ingestion of the transformed records into target system and initial referential integrity processing possibly on both sides.	Provider Institution

Table 2: Summary of the Initial Data Delivery sub-process

6.1.1.1 Syntax Normalization

Syntax normalization aims to convert all data structures relevant for the transformation in a standard form since data transformation tools can only deal with a limited set of standard data structures and thus any non-standard form must be converted to a standard one.

Although most museum systems now employ structured formats and use relational databases, spreadsheets, XML or even RDF, some systems are still dependent upon unstructured storage such as text documents. Even within structured systems there can be issues related to the use of unstructured text fields and misuse of other fields without validation. Automated data transformation, (i.e. transformation of data from one schema to another without loss of meaning or with controlled loss of meaning by a deterministic algorithm based on a mapping definition), is only possible if the data to be transformed is completely structured. Unstructured storage is out of scope for the mapping system and museums will be encouraged to create structured systems for their data. For data issues within structured systems that require

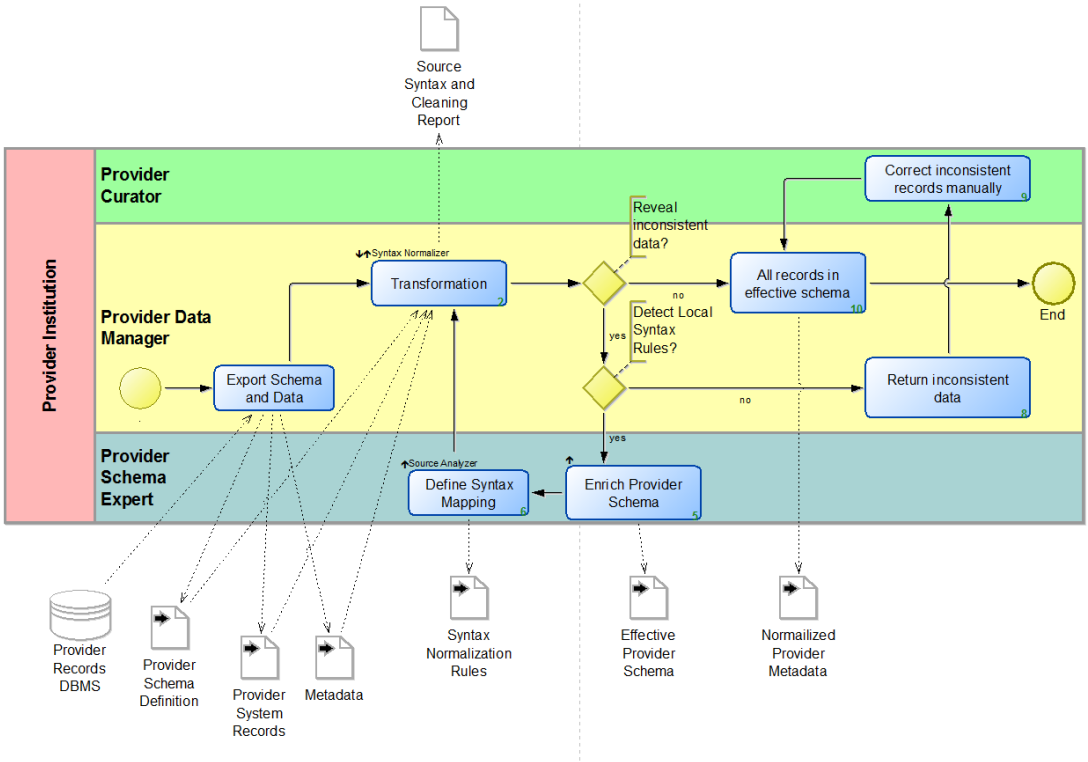

Figure 8: Syntax Normalization sub-process

normalization, the following approaches can be used:

1. The institution resolves these issues in the source database or, if exporting to another format, includes syntax normalization as part of the export process.

2. The mapping system provides a library of syntax normalization routines that can be used by the user.

3. The system provides for new syntax normalization routines that can be created by the institution or their agents. These can potentially become part of a central library or resource for collaborative use.

4. The mapping memory suggests an alternative method for mapping the data that does not capture the full semantics.

Local identifiers may have their own syntactic structure, such as inventory numbers, addresses, bibliographic references, date and time etc. It may not be worthwhile to normalize their internal structure to XML prior to the mapping, but instead use specific and additional scripts for instance generation.

The syntax normalization can be done by a technology expert, possibly the same one dealing with instance generation, in collaboration with the source schema expert. Local syntax rules can be so complicated or even deterministic that it is often more effective to use a set of custom filtering routines, resolving one structural feature at a time, and verifying it with the source schema expert. For instance, if italics are used to tag a particular kind of field, it is better to convert first all italics to XML tags.

This process will reveal inconsistencies and alert the provider to issues that may require attention and need to be resolved. There will be a residual of cases so complicated to describe by rules, that manual rewriting is more effective. This is not a bottleneck, or a reason not to proceed. The important thing is to drastically reduce the number of records to be checked and treated manually. Therefore it is equally important to find diagnostic rules for inconsistent cases, as it is to resolve those that can be formally described. If within a system of syntax normalization some inconsistent cases “slip through” undetected, all records may have to be reviewed manually. That would indeed become a bottleneck. Ultimately no mapping tool can mitigate all internal data management issues and organizations wishing to participate in big data initiatives. Organizations would need to address inconsistencies that perhaps have been historically ignored when data was viewed as simply an internal inventory in a closed system.

After syntax normalization, we expect all data structures relevant for the transformation to be in a standard form. Note that in this step NO approximation of the target schema semantics should be attempted. Rather, it must be an exact representation of the data as understood by the Provider Institution. The CRM is unique in that it represents data as it is understood by the owning organization and does not impose constraints on meaning. There should be no need to manipulate the conceptualization of information to suit the aggregator’s model. The idea is to

capture and bring to the fore the semantics as seen and intended by the provider independently of the aggregator. Table 3 depicts a summary of the tasks presented in Figure 8.

Name	Type	Description	Document		Role	IT Object
			Input	Output		
Export Schema and Data	T	Export schema and data from the Provider Institution	Provider Records DBMS	Provider Schema Definition Provider System Records Metadata	Provider Data Manager	
Transformation	T	Since data transformation tools can only deal with a limited set of standard data structures, any non-standard forms must be converted to a standard one, i.e., RDBMS, XML, RDF OWL or at least spreadsheets. After syntax normalization, we expect all data structures relevant for the transformation to be in a standard form.	Provider Schema Definition Provider System Records Metadata	Source Syntax and Cleaning Report	Provider Data Manager	Syntax Normalizer
Reveal inconsistent data	EG	Check if any inconsistent data has been revealed from the transformation.			Provider Data Manager	
Detect Local Syntax Rules	EG	Check if local syntax rules exist.			Provider Data Manager	
Enrich Provider Schema	T	Enrich the provider schema definition with the formal description of the local syntax rules (Effective Provider Schema).		Effective Provider Schema	Provider Schema Expert	
Define Syntax Mapping	T	Define the correct syntax mapping according to the new schema.		Syntax Normalization Rules	Provider Schema Expert	
Return inconsistent data	T	Return inconsistent data back to the provider in order to manually check them.			Provider Data Manager	
Correct inconsistent records manually	T	Manually review and correct inconsistent data.			Provider Curator	
All records in effective schema	T	After syntax normalization, we expect all data structures relevant for the transformation to be in the effective schema.			Provider Data Manager	

Table 3: Summary of the Syntax Normalization sub-process

6.1.1.2 Mapping Definition

The Mapping definition sub-process includes the specification of the parameters for the data transformation process, such that complete sets of data records can automatically be transformed, manual exception processing notwithstanding. This includes harmonization between multiple providers.

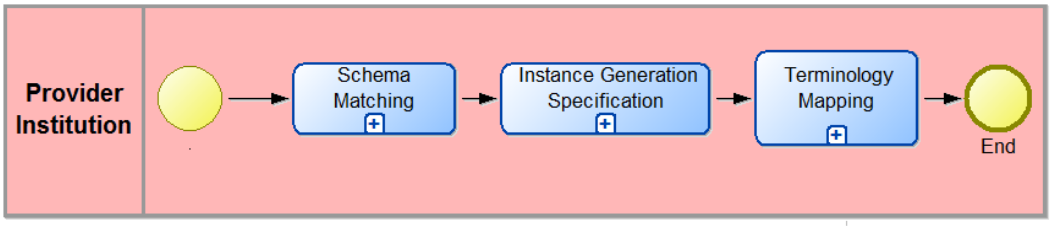

Figure 9: Mapping Definition sub-process

The Mapping Definition sub-process consists of the:

- Schema matching sub-process
- Instance Generation Specification sub-process
- Terminology Mapping sub-process

The Mapping Manager may be responsible for issuing and coordinating these tasks.

Table 4 depicts a summary of the tasks presented in Figure 9.

Name	Type	Description	Organization
Schema Matching	Sub-p	Source schema experts and a target schema expert define a schema matching which is documented in a schema matching definition. In order to do so, all source schema elements must be well understood and mapped to target schema paths.	Provider Institution
Instance Generation Specification Definition	Sub-p	Define the URI generation policies for each instance of a target schema class referred to in the matching	Provider Institution
Terminology Mapping	Sub-p	Define the terminology mappings between source and target terms.	Provider Institution

Table 4: Summary of the Mapping Definition sub-process

6.1.1.2.1 Schema Matching

The provider schema experts together with a target schema expert (e.g., a CIDOC CRM expert) define a schema matching which is documented in a **Schema Matching Definition**. This definition must be human and machine readable and is the ultimate definition of the semantic correctness of the mapping. The collaboration between these experts must be well organized and is the bottle-neck of the data provisioning process.

In order to define a schema matching, all source schema elements must be well understood and mapped to equally well understood target schema paths. Both tasks need two independent tools to visualize source and target schemata and source metadata records. Adequate navigation and viewing techniques must facilitate both overviews and an understanding of the details.

The matching process must lead the user through all source elements in order to make a mapping decision. This may be supported by tools *suggesting mappings* (automated mapping). The automated mapping tools should recalculate their proposals with each new mapping decision. They should make use of "*mapping memories*" of analogous cases collected from the user community. An aggregator may maintain *mapping guidelines* together with provider and user consortia.

As described in 4.2, the schema matching may need to interpret provider or aggregator terminologies in order to re-solve data dependent mappings (where values might determine the mapping). In the Schema Matching Definition, we generally foresee *mapping conditions* that a term in the source record is equal to or unequal to a constant, or a narrower term of a constant. This may be expressed in terms of source or target terminology. In order to resolve these specifications at record transformation time, partial terminology mappings of source and target terminology must exist and may be linked to mapping conditions. The terminology mapping needs to be done only to the degree needed to resolve the conditions of the schema matching. If the provider terminology is hierarchical, the effort can be drastically reduced.

Consider a situation in which a provider has a flat authority with different terms that require different mapping. The potential solutions are as follows:

1. Conditions are provided on an individual basis or, where the majority of terms conform to particular mapping conditions are provided for the exceptions.

2. The terms are grouped into the types that determine their mapping approximating to a single level thesaurus. Conditions can then be applied to those groups.

3. The authority is re-organized into a fully realized thesaurus and conditions applied to branches of the hierarchy.
4. The authority is mapped against the authority/ thesaurus of the aggregator.

As stated above, this mapping may only need to be partial in order to resolve particular automated mapping decisions. Once the mapping is complete then additional semi-automated co-referencing can take place using reasoning across all the aggregator's datasets.

All related tools should take into account the need for incremental mappings after source or target schema definition up-dates, terminology updates and mapping guideline updates and guide the user through relevant changes. These could be highly sophisticated and granular or relate to modifications at a record level triggering a full record update that would include the update.

Some source data or source schema elements may not allow for matching decisions or create a mapping with more general semantics than might be ideal. In situations where the mapping is highly generalized (and not in accordance with the mapping memory) the mapping may be automatically identified as needing an improvement. In some situations, the mapping cannot be made with an improvement to the source by the provider. It must be possible to define filters for these data that run before and at transformation-time and feed back to the provider.

Figure 10Figure 10: Schema Matching sub-process illustrates the Schema Matching sub-process and Table 5: Summary of the Schema Matching Definition sub-process presents a summary of its tasks.

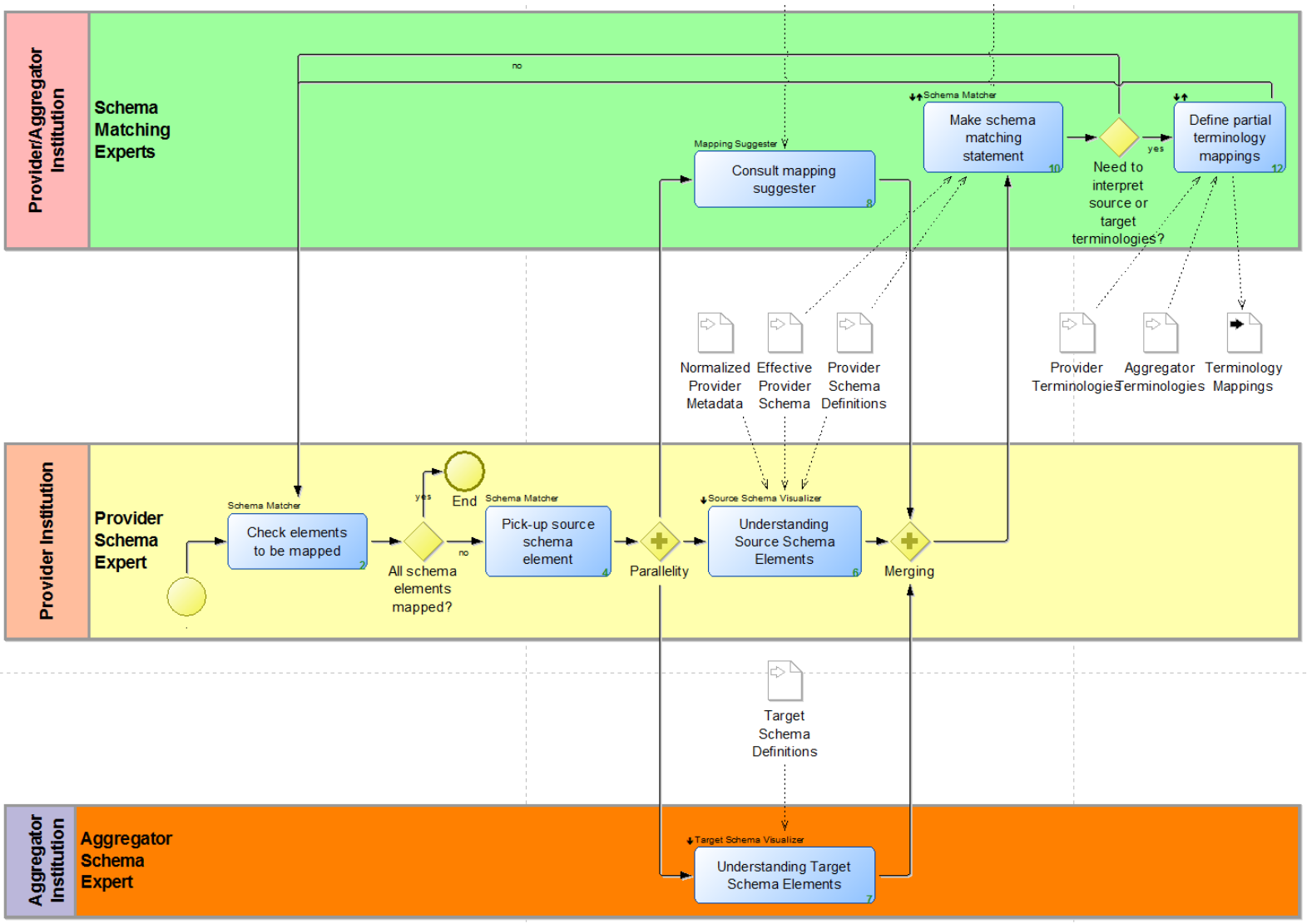

Figure 10: Schema Matching sub-process

Name	Type	Description	Document		Role	IT Object	Organization
			Input	Output			
Check elements to be mapped	T	Check schema matching definition in order to examine if all source schema elements are mapped to target schema paths.			Provider Schema Expert	Schema Matcher	Provider Institution
All elements mapped?	EG	This process continues till all source elements are mapped to a target path.			Provider Schema Expert		Provider Institution
Pick up source schema element	T	Select the source schema path to be mapped.			Provider Schema Expert	Schema Matcher	Provider Institution
Understanding source scheme elements	T	All source schema elements must be well understood. It is important to use tools for the visualization of source schema and source metadata	Normalized Provider Metadata Schema Effective Provider Provider Schema Definitions		Provider Schema Expert	Source Schema Visualizer	Provider Institution
Understanding target schema elements	T	All target schema elements must be well understood. It is important to use tools for the visualization of target schema.	Target Schema Definitions		Target Schema Expert	Target Schema Visualizer	Aggregator Institution

Consult mapping suggester	T	The mapping decision may be supported by tools suggesting mappings. The automated mapping tools should recalculate their proposals with each new mapping decision by some user. They should make use of "mapping memories" of analogous cases collected from the user community. An aggregator may maintain mapping guidelines together with provider and user consortia.	Mapping Memory		Schema Matching Experts	Mapping Suggester	Provider / Aggregator Institution
Make schema matching statement	T	Map a source schema element to a target schema path.	Effective Provider Schema Provider Schema Definitions	Schema Matching Definition	Schema Matching Experts	Schema Matcher	Provider / Aggregator Institution
Need to interpret provider or aggregator terminologies ?	EG	The matching may need to interpret provider or aggregator terminologies in order to re-solve data dependent mappings.					Provider / Aggregator Institution
Define partial terminology mappings	T	In the schema matching definition, we generally foresee mapping conditions that a term in the source record is equal to or unequal to a constant, or a narrower term of a constant. This may be expressed in terms of source or target terminology. In order to resolve these specifications at record transformation time, partial terminology mappings of source and target terminology must exist.	Provider Terminologies Aggregator Terminologies	Terminology Mappings	Schema Matching Experts		Provider / Aggregator Institution

Table 5: Summary of the Schema Matching Definition sub-process

6.1.1.2.2 Instance Generation Specification

After the matching process, an appropriate URI schema must be applied for each target class instance. These URIs are ones defined by a combination of information including the namespace used by the provider, the type of URI (object, terminology, etc.) and the mapping function used. It may also be affected (customized) by particular policies of the provider using the language that best fits the data, typically defined by information managers. Some URIs may be based on third party URI definitions and may require a look-up against appropriate online resources. Changes

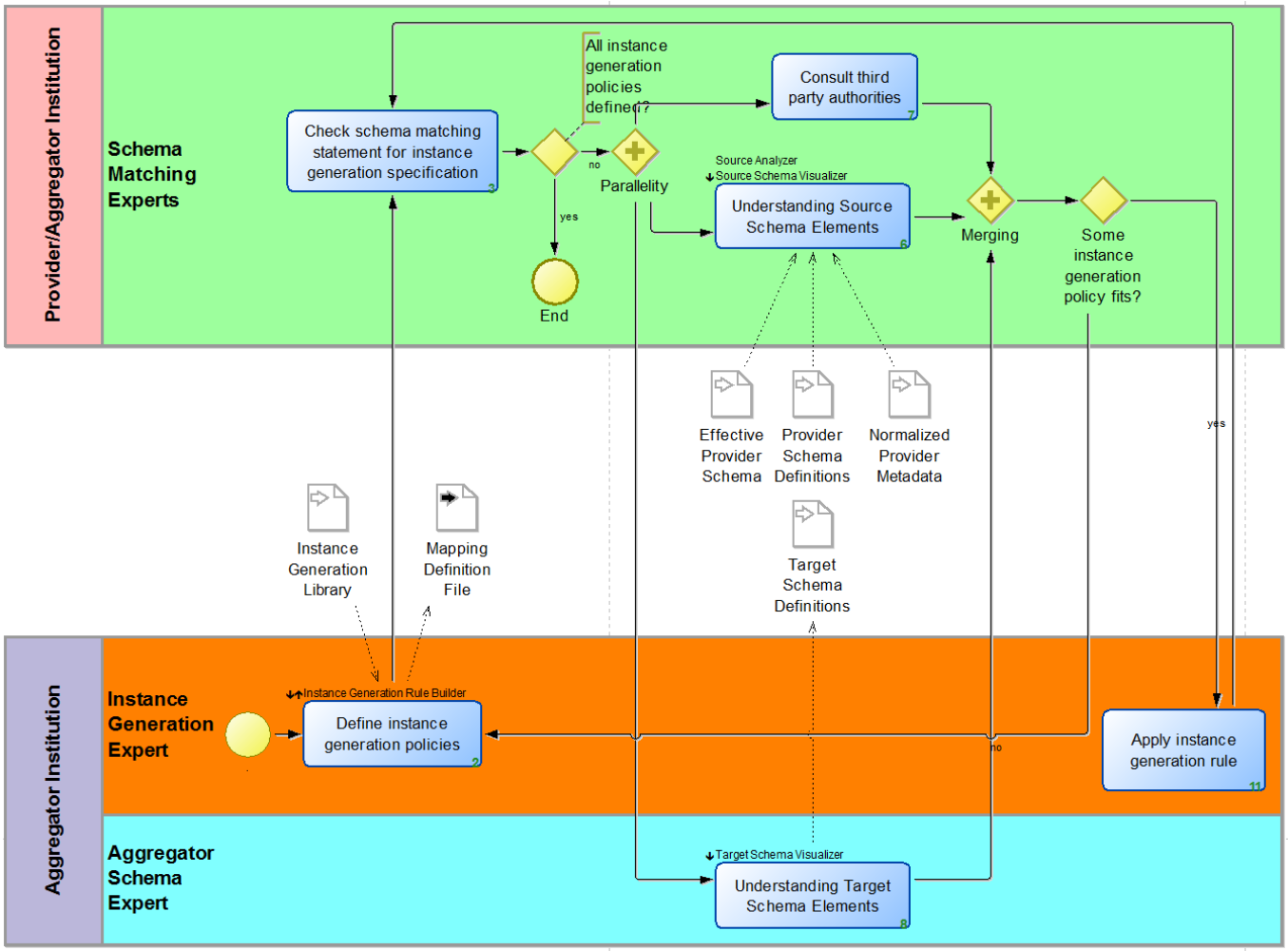

Figure 11: Instance Generation Specification sub-process

in provider instance generation policies may result in changing the definitions without affecting the schema matching. As a result, the application of URIs may be a combination of both automatic and manual steps.

Two situations may exist. Firstly, the provider is interested in representing their data using their own URI schema. For example, they may be using the data internally as part of their own data harmonization strategy and may also wish to create their own reusable interface. In this situation they may wish to use their own URI schema.

Alternatively they may be happy to use the schema applied by the aggregator who may have a URI policy for the one or the other category of items acceptable for the provider. It is likely that the former will become more common and the provider's data should be represented using the providers URIs. If different aggregators use different instance generation policies then co-reference resolution would be needed between URIs that effectively hold the same information. It is ***not*** *the intention* of this model to propose a best practice of URI policies, but to make *the distinctions necessary* to handle consistently and globally resolve the effects of reasonable URI policies.

It is already the case that data providers who publish linked data use the URIs of third parties instead of minting their own URIs. For example, third party URIs are used for vocabularies and ontology support. URI generation policies that use third party resources ("authority records") about persons or places at transformation time should ensure that adequate exception handling is built in, in case of lookup failure.

The execution of URI generation rules may also reveal inconsistent or not normalized data of the provider at transformation time, or before. Inconsistent data filters must be foreseen in the generation rules. The source metadata records may be analyzed before transformation time for such cases. Providers must be informed about inconsistent cases, and given the possibility to run an organized, sustainable process to improve the source data. It may be possible to define preliminary workarounds to maintain the submission process, i.e. characteristic data patterns replacing dirty data that can later be recognized and updated at aggregator side without resubmission. Otherwise, inconsistent records are held back until they are updated.

Changes of instance generation policies of the aggregator may result in the need to update the URI generation rules. Changes of naming and identifier policies at the provider side may also make a redefinition of URI generation rules necessary. It must be possible to do that without affecting the schema matching definition file.

Table 6: Summary of the Instance Generation Specification sub-process depicts a summary of the tasks presented in Figure 11.

Name	Type	Description	Document		Role	IT Object	Organization
			Input	Output			
Define instance generation policies	T	The URI generation policies for each instance of a target schema class referred to in the matching must be defined, such as for persons, objects, events, place, and formats of time. The URI generation policies can be introduced in an abstract form as rules or references to code signatures implementing specific rules.	Instance Generation Library	Mapping Definition File	Instance Generation Expert	Instance Generation Rule Builder	Aggregator Institution
Check schema matching statement for instance generation specification	T	Examine each instance of a target schema class to apply a URI generation rule.			Schema Matching Experts		Provider / Aggregator Institution
All instance generation policies defined?	EG	Check each instance of the target schema class, if a URI generation rule is applied.			Schema Matching Experts		Provider / Aggregator Institution
Understanding source schema elements	T	All source schema elements must be well understood. It is important to use tools for the visualization of source schema and source metadata records.	Effective Provider Schema Provider Schema Definitions Normalized Provider Metadata		Schema Matching Experts	Source Analyzer Source Schema Visualizer	Provider / Aggregator Institution
Consult third party authorities	T	Some URI generation policies may include look-up of on-line resources (“authority records”) about persons or places.			Schema Matching Experts		Provider / Aggregator Institution
Understanding Target Schema	T	All target schema elements must be well understood. It is important to use tools for the visualization of target schema	Target Schema Definitions		Target Schema Expert	Target Schema Visualizer	Aggregator Institution
Some instance generation policy fits?	EG	Some instance generation policy fits?			Schema Matching Experts		Provider / Aggregator Institution

Apply instance generation rule	T	Apply the instance generation rule, in case a instance generation policy fits.			Instance Generation Expert		Aggregator Institution

Table 6: Summary of the Instance Generation Specification sub-process

6.1.1.2.3 Terminology Mapping

Terminology mapping can be a huge task. Providers may use anything from intuitive lists of uncontrolled terms up to highly structured third party thesauri. However, most of the provider terminology is very specialized and more important as information element, when metadata records are displayed, than as search term in the target system. As long as the terms are in the same natural language, most terms can just be copied from the source records into the transformed target records. If they are in other languages, aggregators may choose to translate terms, possibly preserving also the provider terms. It may be useful to associate provider terms with broader terms of some standard terminology the aggregator employs as search terms. Since all this can happen even after metadata record transformation, it does not affect the mapping process itself.

In this model, we are only interested in the consistency of the mapping process when the choice of a target class or property depends on a term. For that sake, we can extract from the schema matching definition the terms appearing in mapping conditions. We distinguish two cases:

- ***equality/inequality condition:*** These terms (constants) must be taken from the provider terminology and no action is needed.
- ***broader term condition:***
  1. If the constant term is given in the provider terminology, the narrower term hierarchy for each constant term is used if it exists, otherwise it must be “invented”. The latter is one case of terminology mapping.

  2. If the constant term is given in the aggregator terminology, for each constant term the narrower terms in the source terminology must be identified. This is the second case of terminology mapping.

In any case, the aggregator terminology should have a thesaurus structure, albeit a small vocabulary of high-level terms. Sometimes it may be more effective to merge provider terms with aggregator terms, i.e., replace equivalent terms and insert all other provider terms as narrower terms of aggregator terms.

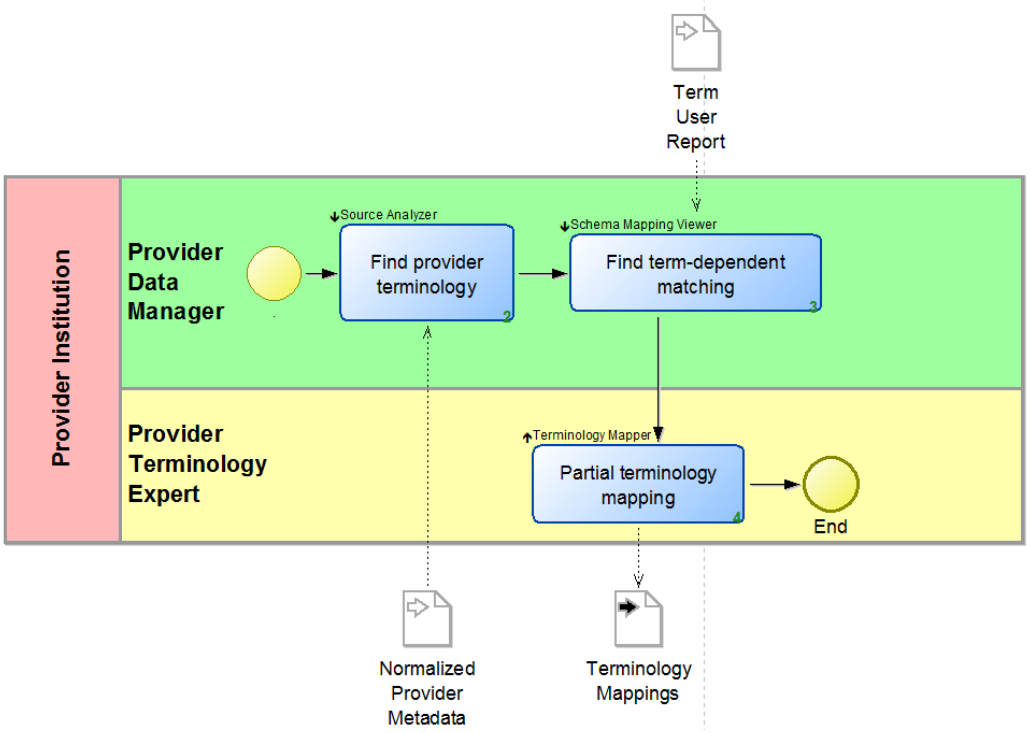

Figure 12: Terminology Mapping sub-process

In this case, the term values used to execute the schema matching conditions of the provider terminology should be replaced by the updated aggregator terminology before transforming the respective records in the Schema Matching Definition. This will allow for better controlling the mutual consistency of mappings between different providers. Notwithstanding, the original provider terminology could, possibly should, be added to the source records and be carried over to the target records in separate fields/properties.

The terminology mapping may reveal inconsistent data of the provider, such as spelling errors or unauthorized terms. Inconsistent data filters must be foreseen for terminology. The source metadata records should be analyzed before transformation time for such cases. Providers must be informed about inconsistent data cases, and given the possibility to run an organized, sustainable process to improve the source data.

It may be possible to define preliminary workarounds to maintain the submission process, i.e. characteristic data patterns replacing dirty data that can later be recognized and updated at aggregator side without resubmission. Otherwise, “dirty records” are held back until they are updated.

Mapping identifiers of persons and places to those used at the aggregator system is in general ineffective due to the large number of such identifiers, most of which are only known at local level. It is more effective to update the provider with identifiers (URIs) of persons and places referred to by more than one provider (or third party

authority, such as viaf.org). After these steps, metadata records are ready for transformation.

Table 7 depicts a summary of the tasks presented in Figure 12.

Name	Type	Description	Document		Role	IT Object	Organization
			Input	Output			
Find provider terminology	T	The provider data manager uses the source analyzer tool in order to extract the terms appearing in the source schema.	Normalized Provider Metadata		Provider Data Manager	Source Analyzer	Provider Institution
Find term-dependent matching	T	Define a partial terminology mapping of source and target terminology..		Terminology Mappings	Provider Terminology Expert	Terminology Expert	Provider Institution
Partial Terminology Mapping	T	Check each instance of the target schema class, if a URI generation rule is applied.			Schema Matching Experts		Provider / Aggregator Institution

Table 7: Summary of the Terminology Mapping sub-process

6.1.1.3 Metadata Transformation

Once the mapping definition has been finalized (and all syntax errors resolved) the data and mapping information needs to be submitted, transformed and stored in the aggregators system.

The mapping manager will be informed of the submission to initiate the transformation process, provide final validation and store both the raw data (optional) and transformed data into the target system. The submitted metadata records must be identified with a unique identifier, checksum and modify date-time. The submission management must be able to recognize any change of a source record by the metadata record.

Transformation

The transformation process itself may run completely automatically. However, it is possible that further issues not realized by the provider will materialize. These are issues that the aggregator may be able to resolve on a temporary or permanent basis

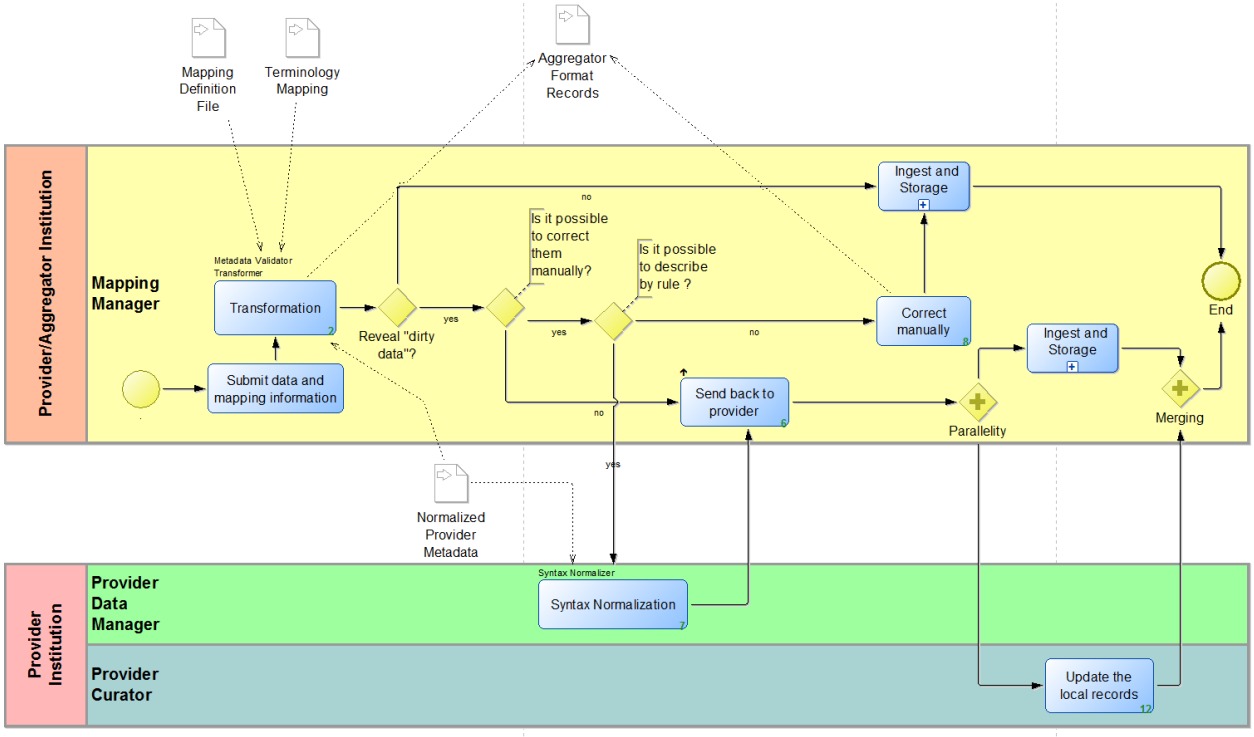

Figure 13: Metadata Transformation sub-process

but in any event may require that records are referred back to the provider for further analysis and resolution. It is possible (given sufficient trusted expertise) that personnel under the aggregator mapping manager can take manually correct issues in an interactive process. The result is a set of valid target records.

Instance Matching

The instance generation algorithm of the automatic data transformation process may employ an initial instance matching process in order to reuse existing URIs in the target system. This also holds for third party authority systems with URIs that the aggregator uses as reference (such as viaf.org). In any case, such matching should be reported back to the provider for potential internal use of such URIs. The provider should not replace local identifiers of items under his authority of knowledge aggregators system. Alternative URIs identified by the aggregator should only be used in addition to established local data for which the provider can verify the referred thing, and aggregation should not be used to homogenize provider data. **Table 8** below depicts a summary of the tasks presented in Figure 13.

Name	Type	Description	Document		Role	IT Object	Organization
			Input	Output			
Submit data and mapping information	T	Once the mapping definition has been finalized, the data and mapping information needs to be submitted.			Mapping Manager		Provider / Aggregator Institution
Transformation	T	Transform source metadata to target records, ready to be ingested to the target system	Normalized Provider Metadata Mapping Definition File Terminology Mapping	Aggregator Format Records	Mapping Manager	Metadata Validating Transformer	Provider / Aggregator Institution
Reveal "dirty data"?	EG	Transformation process may reveal inconsistent data that need to be mended.			Mapping Manager		Provider / Aggregator Institution
Is it possible to correct them manually?	EG	Is it possible to correct inconsistent data manually?			Mapping Manager		Provider / Aggregator Institution
Is it possible to describe by rule?	EG	Check if local syntax rules exist.			Mapping Manager		Provider / Aggregator Institution

Send back to provider	T	Sort out dirty records and send them back to the provider for processing.			Mapping Manager		Provider / Aggregator Institution
Syntax Normalization	T	Convert data to the structural format described by the rules.			Provider Data Manager	Syntax Normalizer	Provider Institution
Correct manually	T	If possible and given sufficient, trusted expertise, experts under the control of the mapping manager may correct some of them manually in an interactive process. The result is a set of valid target records.		Aggregator Format Records	Mapping Manager		Provider / Aggregator Institution
Ingest and Storage	Sub-p	The transformed records are ingested into the target system.			Mapping Manager		Provider / Aggregator Institution
Update the local records	T	The provider should better update his records.			Provider Curator		Provider Institution

Table 8: Metadata Transoformation sub-process

6.1.1.3.1 Ingest and Storage

Once records are transformed, an automated translation for source terms using a terminology map may follow. The transformed records will then, be ingested into the target system.

An Ingest Manager should also store all source metadata records for the transformed information. This is considered very good practice and supports transparency important for academic projects. The Ingest Manager must preserve a link to the identity and version of the source record it is derived from. Some aggregators additionally provide source data to users of the target system as part of query results (e.g., the German Digital Library).

In case a new version of a transformed source record is ingested in the target system, the target record representing the previous version must be removed for the purposes of canonical searching. Some aggregators will keep previous versions for historical and academic purposes but these versions should be separate and correctly described to avoid any confusion. However many aggregators will have a deletion policy and it is advisable that providers keep old versions. Otherwise the provider and aggregator may agree terms to manage and preserve old versions as an

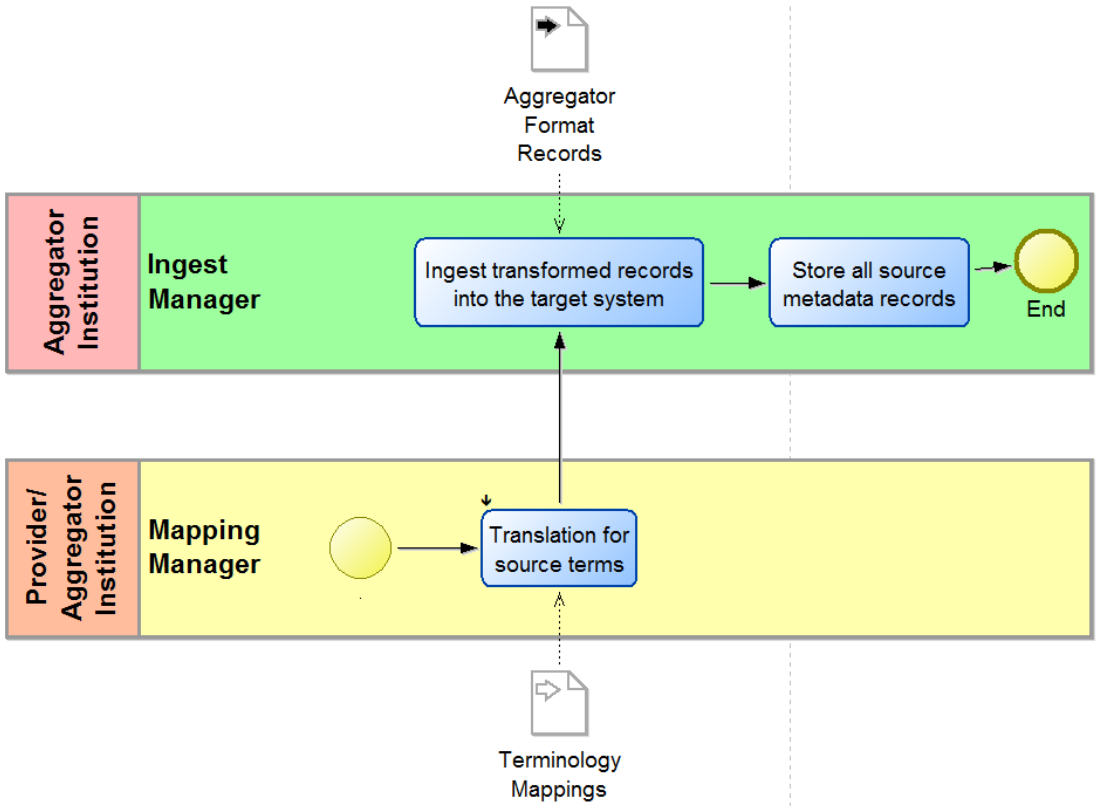

Figure 14:Ingest and Storage sub-process

additional service. In this case providers must make sure that the data is recoverable in the same format that it was submitted in addition to the aggregator's model.

Table 9: Summary of the Ingest and Storage sub-process depicts a summary of the tasks presented in Figure 14.

Name	Type	Description	Input Document	Role	Organization
Translation for source terms	T	An automated translation for source terms using a terminology map may follow.	Terminology Mappings	Mapping Manager	Provider / Aggregator Institution
Ingest transformed records into the target system	T	The transformed records will be ingested into the target system.	Aggregator Format Records	Ingest Manager	Aggregator Institution
Store all source metadata records	T	An aggregator should store all source metadata records which are going to be transformed, or which are transformed and have been ingested to the target system. The aggregator must preserve the link to the identity and version of the source record it is derived from. Some aggregators return also fitting source records in query results (e.g., the German Digital Library). Source records may be syntactically normalized for that purpose.		Ingest Manager	Aggregator Institution

Table 9: Summary of the Ingest and Storage sub-process

6.1.2 Update Processing

The mapping manager must monitor all changes that may affect the consistency of provider and aggregator data.

Those changes are:

1. New source records
2. Source record updates (new versions)
3. Source schema changes
4. Provider changes identifier policy (for people, objects, events, places, time) and updates his records
5. Provider changes terminology data (terms or authority) and updates his records
6. Provider changes terminology structure (broader term links etc.)
7. Target schema changes
8. Aggregator changes URI policy
9. Aggregator changes terminology (terms or authority)
10. Aggregator or user consortium changes mapping guidelines
11. Source-target terminology mapping changes

The changes of number 1 and 2 require running the complete metadata transfer with the changed or new source records but using the existing mapping definition.

Changes of kind 3 require updating the schema matching definition in the mapping file, resubmission of all source records affected, transformation and ingestion, replacing the target records transformed from the previous version of these source records.

Changes of kind 4 require updating the URI generation specification in the mapping file, resubmission of all source records affected, transformation and ingestion, replacing the target records transformed from the previous version of these source records.

Changes of kind 5 require updating the terminology mapping, resubmission of all source records affected, transformation and ingestion, replacing the target records transformed from the previous version of these source records

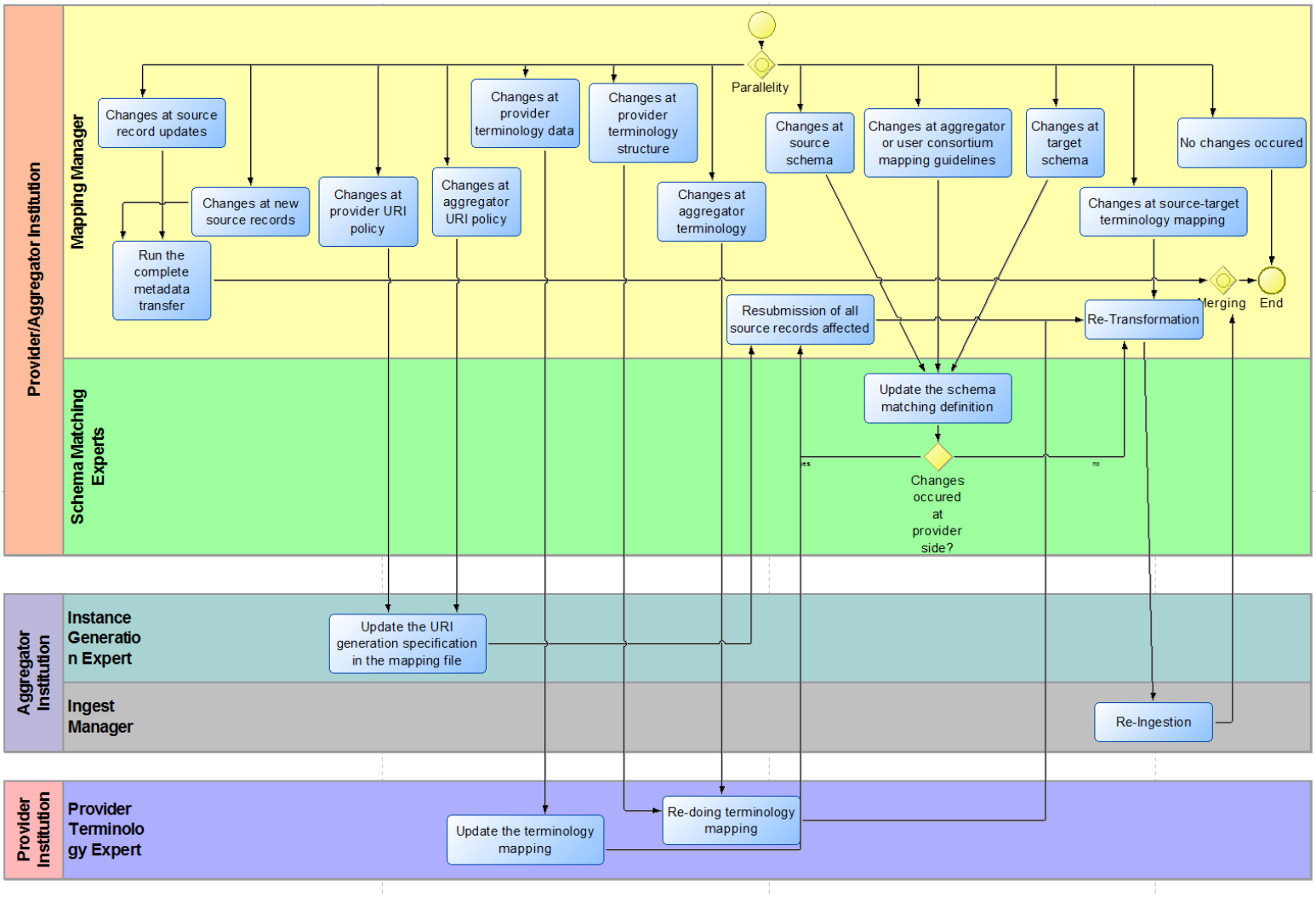

Figure 15: Update Processing sub-process

The changes of kind 6 and 9 require a rework of the terminology mapping, retransformation and re-ingestion of all (stored) source records already transferred which refer to the respective terms.

Changes of kind 7 and 10 require updating the schema matching definition, retransformation and re-ingestion of all (stored) source records already transferred which refer to the respective terms.

Changes of kind 8 require updating the URI generation specification in the mapping file, retransformation and re-ingestion of all (stored) source records already transferred which refer to the respective terms.

The changes of kind 11 require retransformation and re-ingestion of all (stored) source records already transferred which refer to the respective terms.

Table 10 depicts a summary of the tasks presented in Figure 15.

Name	Type	Description	Role	Organization
Changes at source record updates	T	Changes at source record updates (new versions)	Mapping Manager	Provider / Aggregator Institution
Changes at new source records	T	New source records	Mapping Manager	Provider / Aggregator Institution
Changes at provider URI policy	T	Provider changes identifier policy (for people, objects, events, places, time) and updates his records	Mapping Manager	Provider / Aggregator Institution
Changes at aggregator URI policy	T	Aggregator changes URI policy	Mapping Manager	Provider / Aggregator Institution

Changes at provider terminology structure	T	Provider changes terminology structure (broader term links etc.)	Mapping Manager	Provider / Aggregator Institution
Changes at provider terminology data	T	Provider changes terminology data (terms or authority) and updates his records	Mapping Manager	Provider / Aggregator Institution
Changes at aggregator terminology	T	Aggregator changes terminology (terms or authority)	Mapping Manager	Provider / Aggregator Institution
Changes at source schema	T	Changes at source schema	Mapping Manager	Provider / Aggregator Institution
Changes at target schema	T	Target schema changes	Mapping Manager	Provider / Aggregator Institution
Changes at aggregator or user consortium mapping guidelines	T	Aggregator or user consortium changes mapping guidelines	Mapping Manager	Provider / Aggregator Institution
Changes at source-target terminology mapping	T	Source-target terminology mapping changes	Mapping Manager	Provider / Aggregator Institution
No changes occured	T	No changes occured	Mapping Manager	Provider / Aggregator Institution
Run the complete metadata transfer	T	Run the complete metadata transfer with the changed or new source records but using the existing mapping definition.	Mapping Manager	Provider / Aggregator Institution
Resubmission of all source records affected	T	Resubmission of the affected source records	Mapping Manager	Provider / Aggregator Institution

Retransformation	T	Retransformation of the affected records	Mapping Manager	Provider / Aggregator Institution
Re-ingestion of all (stored) source records	T	Re-ingestion of all (stored) affected records.	Ingest Manager	Aggregator Institution
Re-doing terminology mapping	T	Redoing the terminology mapping	Provider Terminology Expert	Provider Institution
Update the schema matching definition	T	Update the schema matching definition	Schema Matching Experts	Provider / Aggregator Institution
Update the URI generation specification in the mapping file	T	Update the URI generation specification in the mapping file.	Instance Generation Expert	Aggregator Institution
Update the terminology mapping	T	Update the terminology mapping	Provider Terminology Expert	Provider Institution

Table 10: Summary of the Update Processing sub-process

7 Services and Software components

This section describes in detail the software components foreseen by this model. The intention of this model is only to define interoperable interfaces between the components, such that an effective monitoring and workflow control system can interact with the components and combinations from arbitrary providers in arbitrary technologies can interact to enable the whole process with all its details. It should further enable rich enough variations of particular workflows and extension of functionality. Each component may be implemented with different levels of sophistication, from simple commands to fancy graphic manipulations. Some components may only support limited functionality, e.g., transformation from XML to XML. In such cases, the workflow system should be able to plug in on demand alternative components for other formats, such as E-R to XML, E-R to RDF, XML to RDF etc.

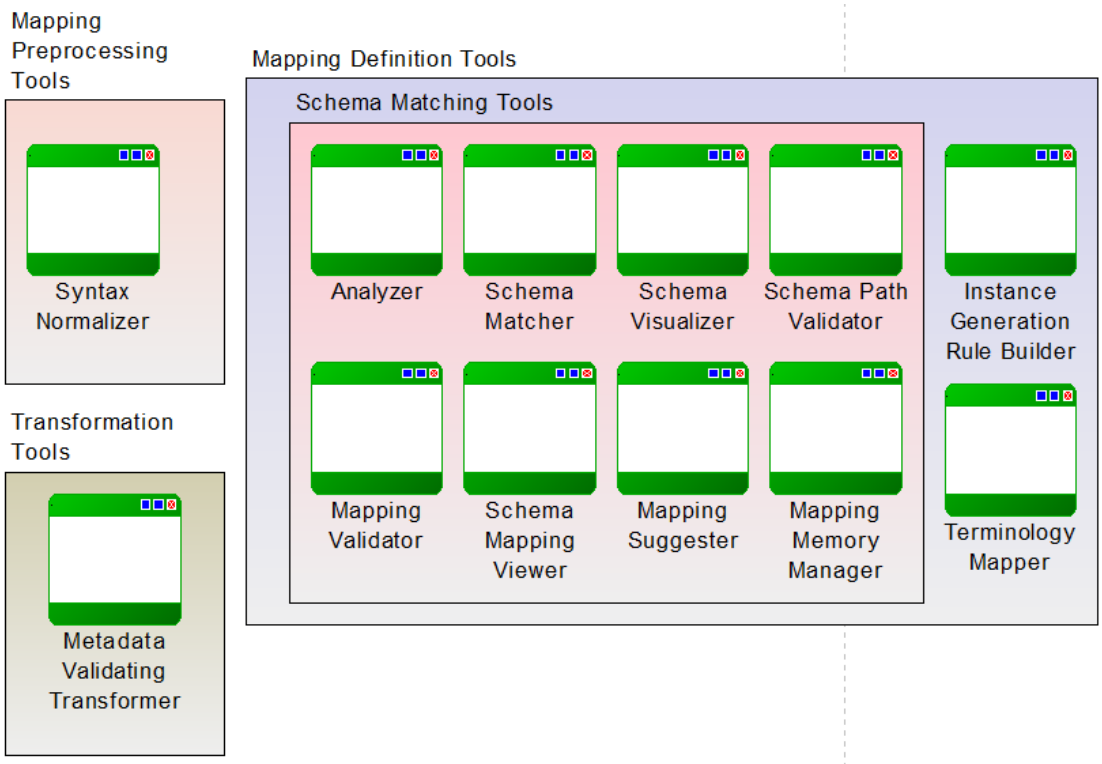

Figure 16: Services and S/W components

Figure 16 Illustrates the IT objects that assist the data provisioning process. We do not regard IT objects as self-contained and opposed to user processes, but IT objects are regarded as being part of the user processes replacing or supporting manual work. As depicted in Figure 16: Services and S/W components, we group the IT components into four different categories:

- Mapping Preprocessing Tools
- Mapping Definition Tools
- Transformation Tools

In the following sub sections we describe in detail the components of each category.

This section will also be extended by exact interface definitions.

7.1 Mapping Preprocessing Tools

The Mapping Preprocessing Tools consists of the IT objects that assist the experts at the preprocessing stage for the normalization of data.

We define the following mapping preprocessing tools:

Syntax Normalizer

Description

The Syntax Normalizer consists of two different functionalities. The first step of the syntax normalization is to convert any data structure that has no standard format, like csv, excel to a standard one, like XML, JSON, RDF, OWL.

Then, the syntax normalization follows on the field level, which aims to detect and correct inconsistencies in the text format of field values. Sometimes this involves revealing hidden schema structure “encoded” in fields, such as names separated by semicolons, in which case the syntax normalization must be able to affect local changes in the schema structure to capture the hidden structures. Other syntax normalizations may include combining the contents of several fields into one, or transforming identifier values into URLs to access resources such as media files.

The Syntax Normalizer should work as interactively as possible so that data specialists can build and test the normalizations on live data. It should consist of a validation/selection phase followed by a manipulation phase, and the interaction should be a process of progressive refinement involving repeated filtering into valid/invalid values until the number of invalid is minimal. It should also contain a building library of reusable processes.

7.2 Mapping Definition Tools

The Mapping Definition Tools consists of a set of various components able to transform content and metadata, in various, heterogeneous formats, to a normalized data model such as the CIDOC CRM.
The Mapping Definition Tools is further subdivided to the Schema Matching Tools that support the schema matching process. Takes as input two ontologies and determines the alignment result between entities of the input ontologies.

We define the following mapping definition tools:

Analyzer

Description

The Analyzer dissects the source and target data, providing a useful set of statistics for each field and a comprehensive analysis of schema structure and record content, including all unique values, value histograms, and statistic visualizations. During the schema matching process, the target records are not yet known, but we still may have the need to analyze the target schema, in order to create more accurate mappings.

More precisely this component should at least:

- Count the number of tags and visualize each tag
- Compile the complete unique value lists and provide frequencies for each value
- Provide visual representations of the value lists with different sorting functionalities or even randomly
- Provide information about correlated tags
- Count the number of parent tags where an element occurs (in case of XML)
- Count the number of references of a specific table by other tables (in case of relation databases)
- Count the number of properties that have a specific class as range (in case of RDF)
- Count the number of properties that use instances of a specific class as range (in case of RDF)

Schema Visualizer

Description

Tools to visualize source and target schema definition. This component should be developed to deal with different formats, including JSON, XML exports, RDF and potentially other formats. In this respect, it is should be open to the development of additional modules designed to support other formats of data. Adequate navigation and viewing techniques must allow for overviews and understanding of details. It should incorporate and display in context the scope notes, definitions, examples etc. that are provided with the schema and also provide visual representations of the schemas within their hierarchical context. The hierarchical view should be expandable and collapsible at each level and overall (expand all).

The schema should be presented in a way that is easy to navigate and should expose relevant associated information when viewing any particular property. For example, it should be easy to isolate and browse information on related properties or sub-properties. The mapping memory might include information about related properties that are often confused and regularly misused. This information could be made available so that the user is alerted to subtle differences in the schema.

Schema Path Validator

Description

Tools to validate the source and target schema definition. This IT object should be able to deal with different formats, including JSON, XML exports, RDF, OWL and potentially other formats. In this respect, it is should be open to the development of additional modules designed to support other formats of data. The Schema Validator traverses the schema and "proposes" to the experts a list of valid paths. It should also be able to support the validation of a specific path, or triples.

Schema Matcher

Description

The Schema Matcher allows experts to loop through all source elements in order to make a mapping decision. It may be supported by tools suggesting mappings and should recalculate their proposals. Experts should be able to define meaningful matching definitions that preserve the semantics of the source records by adding the appropriate information. The Schema Matcher should allow experts to add information about the mapping and also declare and upload files that are going to be used as source or target schemata, example source or target records and a generator policy file. The uploaded files are then parsed, analyzed and used by the corresponding components.

Schema Mapping Viewer

Description

Tools to visualize the schema matching definition file. Adequate navigation and viewing techniques must allow for overviews and understanding of details. Different interaction views should be supported in order to provide to the experts a quick and clear overview of the defined mappings.

Mapping Validator

Description

Tools to validate the schema matching definition as a whole. It should be able to validate both the source and the targets paths. Similar to the Schema Path Validator, the Mapping Validator should also deal with different formats and potentially facilitate the development of additional modules designed to support other formats of data. The Mapping Validator should "capture" inconsistences created due to previous mappings, for instance, the case of two disjoint classes in RDF. Moreover, the Mapping Validator is responsible for analyzing the Mapping Definition, in order to extract the terms that appear in mapping conditions.

Mapping Suggester

Description

The Mapping Suggester facilitates the schema matching process by suggesting mappings to the experts. These suggestions make use of "mapping memories" of similar cases collected from the user community and are recalculated with each new mapping decision. With the creation of a new mapping file, the Mapping Suggester runs a schema matching with the source schema provided at the new mapping file and the existing mapped source schemata in the mapping memory. The correspondences/crosswalks found during the schema matching are used by the Mapping Suggester to suggest mappings to the user. The expert can either accept or reject the suggestion.

Instance Generation Rule Builder

Description

The Instance Generation Rule Builder facilitates the IT experts to define instance generation rules for each independent node. The component uses the Instance Generator Library templates for specifying how a URI or a label will be created. The generator templates are defined separately and linked to the actual mapping, to be used by the generators for producing the actual identifier or label.

Mapping Memory Manager (3M)

Description

Mapping Memory Manager is a tool for managing mapping definition files. It provides a number of administrative actions that assist the experts to manage their mapping definition files. The component allows the users to create new mapping files. The newly created file can be further edited with the Schema Matcher and the Instance Generation Rule Builder. Since the number of mappings can become quite big, several actions (such as searching, filtering and sorting) should be provided. Moreover, experts should be able to export their mappings for off-line use, import mapping files, create versions of a specific mapping file, delete it and also make a copy.

Terminology Mapper

Description

Terminology Mappings are expressions of exact or inexact (broader/narrower) equivalence between terms from different vocabularies. Terms being mapped are those arising from use in practice in archives, museums and library collection registration systems, and they are typically mapped to URIs representing concepts in shared vocabularies.

In this context, the Terminology Mapper should be used to primarily map the terms that appear directly or indirectly in mapping conditions of a Schema Matching Definition. In such a mapping condition, a term in the source record is equal to or unequal to a constant, or a narrower term of a constant. As a result, this particular mapping may involve a kind of categorization of terms rather than detailed equivalence mapping with a large vocabulary. This may be expressed in terms of source or target terminology.

The terminology mappings may be expressed in SKOS RDF or also in an XML format that will cover the concepts of exact/inexact term and broader/narrower term.

7.3 Transformation Tools

The Transformation Tools consists of the IT objects that assist the experts to transform their data from one standard format to another one, with the use of the mapping definition.

We define the following transformation tools:

Metadata Validating Transformer

Description

The Metadata Validating Transformer realizes the transformation of the source records to the target format. The tool takes as input the provider data and the mapping definition. The Metadata Validator also keeps a Source to Target Association Table for the produced “values”. This means that whenever a new “value” is created (URI, UUID, or literal) the exact XPATH from the source file is associated with the generated value. The contents of this association table should be exported in order to be exploited from the user as a quick overview of the generated values for particular resources of the source.

8 References

[1] ADONIS-BOC Group, [Online]. Available: http://www.adonis-community.com/. [Accessed 2014]

[2] CIDOC-CRM, “The official release of CIDOC-CRM,” [Online]. Available: http://www.cidoc-crm.org/docs/cidoc_crm_version_5.0.4.pdf. [Accessed 2014].