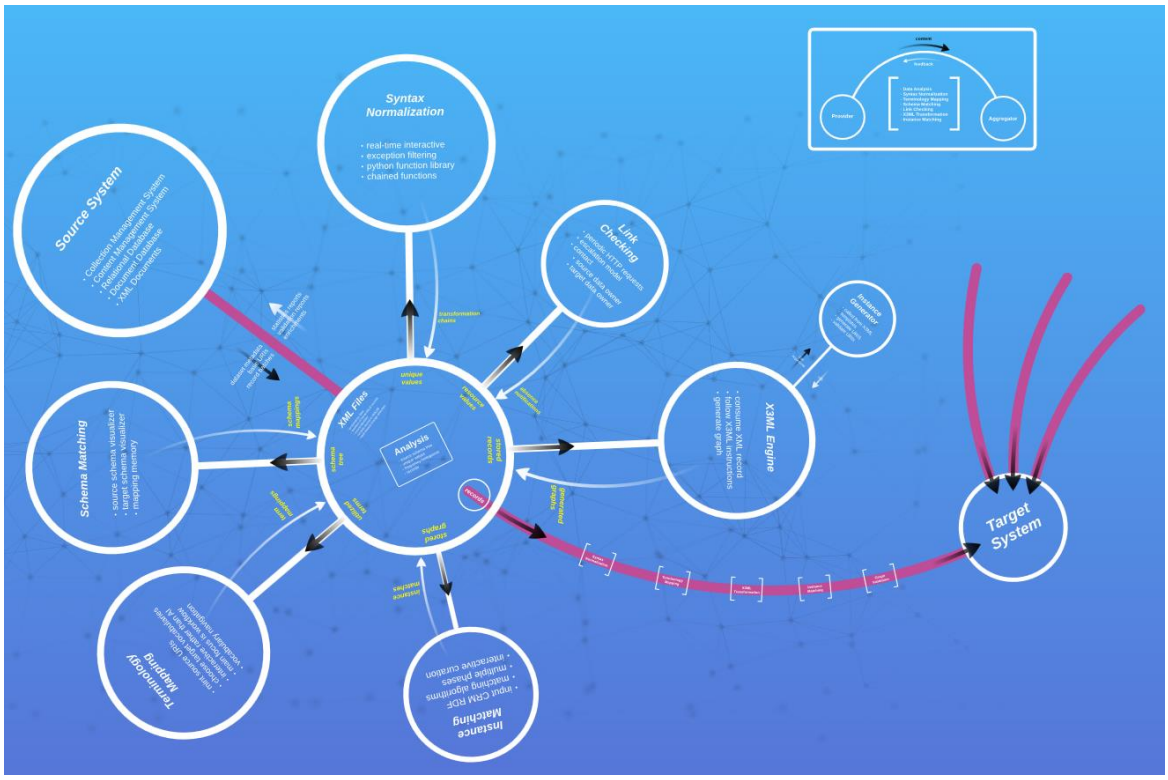
The Relevance of Aggregators: Proposal for a new Data Provision Reference Model - Synergy

Authors: Dominic Oldman, Chrysoula Bekiari, Martin Doerr, Gerald de Jong, Barry Norton, Thomas Wikman

Conference Paper: CIDOC - International Documentation Committee of ICOM: Strategies and Policies in Documentation

Session A/3, Strategies and Policies in Documentation
Wednesday, 10.9.2014



Introduction

Despite the success of museums in becoming more inclusive and accessible to a wider range of visitors, and the associated growth in visitor numbers, the significant and relevance of museums continues to be a concern. Securing the long term role of museums in society is dependent on the strength of relationship it builds with many different groups and the richness of message that it conveys in different mediums, including digital infrastructures. In recent years the question of relevance has been addressed in a digital setting by attempting to aggregate the data of different cultural institutions to create integrated resources. These initiatives, for example, Europeana, attempt to address the problem of isolated data silos by developing centralised integrated repositories in an attempt to encourage multi-faceted data services not achievable by individual institutions. In effect, the integration of different cultural heritage datasets creates a more valuable and useful resource.

However, just as there are concerns about long term sustainability and the intellectual strength of relationships with physical visitors, centralised data repositories have also been criticised for providing weak and superficial representations incapable of building long term relationships with data providers or providing sustainable platforms for digital applications (Janes, 2009). In failing to address key structural design issues, data aggregators can persuade cultural heritage institutions neither to invest sufficient resources, nor to align their internal digital strategies with them; a realistic model of sustainability thereby remains out of reach. This situation perpetuates a weak and fragmented landscape of disparate cultural digital services. This paper identifies some key problems with current data aggregation and reuse services. It also proposes a new data provisioning reference model called, 'Synergy' which is aimed at resolving these fundamental issues and specifying requirements that treat data providers as integral parts of the data provision process and ensures tangible benefits for both data providers and data users alike.

The, 'Synergy Reference Model of Data Provision and Aggregation'[1] is currently a draft proposal created to generate discussion and encourage input from the community to contribute towards a final specification. It is relevant to the aggregation of all cultural datasets, not just those provided by museums, and provides the necessary infrastructure for sustainable provisioning of data from museums, archives and libraries, as well as many other specialist datasets.

Currently it draws upon the work of the CIDOC CRM Special Interest Group (CRM SIG), a working group of CIDOC, the International Committee for Documentation of the International Council of Museums (ICOM). The Group is responsible for developing the CIDOC Conceptual Reference Model (Doerr, 2003; Crofts et al., 2011) and has provided advice and consultancy on cultural heritage data integration for over 16 years. The draft reference model has been motivated by the rapid growth of cultural heritage digital data making the need for a best practice framework more urgent that ever before.

The Position Discussed of the CRM SIG

The following positions are based on a history of over 20 years of data integration projects that have attracted large amounts of funding, but have so far failed to establish a sustainable integrated cultural data platform. The first position is based on a discovery made in the 1990s after the failure

[1] http://www.cidoc-crm.org/docs/SRM_v0.1.pdf

of significant initiatives in both Europe and the United States, and can be summarised by the following quote:

"Those engaged in the somewhat arcane task of developing data value standards for museums, especially the companies that delivered collections management software, have long had to re-present the data, re-encode it, in order for it to do the jobs that museums want it to perform It's still essentially impossible to bring data from existing museum automation systems into a common view for use for noncollections management purposes as the experience of the Museum Educational Site Licensing (MESL) and RAMA (Remote Access to Museum Archives) projects have demonstrated. Soon most museums will face the equally important question of how they can afford to re-use their own multimedia data in new products, and they will find that the standards we have promoted in the past are inadequate to the task."(Bearman, 2010, p.49)

"Increasingly it seems that we should have concerned ourselves with the relationships (creating, selling, designing, using, critiquing) between the objects and the proper nouns on which we lavished so much attention because as we examine the queries being put to us by our publics, it is obvious that each user community needs to know about quite different relations (and, as argued earlier, the nouns could be "controlled" without imposing conformity anyway)." (Bearman, 2010, p.53)

These quotes come from a 1995 essay (reproduced in 2010), acknowledging that the use of fixed field/value models for integration could never provide a satisfactory solution for cultural data integration. Fixed field models do not reflect the variability of the different data sources and ultimately misrepresent data, making data connections unreliable, and of limited use. With no long term value for the data provider, and with limited search and retrieval services for the user, such systems are destined to gradually fade away. Cultural heritage data aggregation needs to retain the meaning and context of source systems otherwise users will be perpetually beholden to refer back to the original source, defeating the purpose of a modern central resource or reuse service.

The CIDOC CRM ontology, initiated in 1996, was designed to resolve the issue of using fixed field models to integrate data. It replaced the old CIDOC Entity-Relationship model of cultural heritage data (a model original designed by the Smithsonian institute and realised in relational databases) with a new object-oriented model. Instead of simply adding growing numbers of entities and fields to an already bloated model this new approach concentrated on a semantic framework using relationships and events to provide a maintainable model in which all datasets, regardless of their difference and variability. Such a model can be digitally represented without imposing any restrictions on the data or terminologies used. (Doerr and Crofts, 1998)

Once one accepts the premise that data aggregations should attempt to retain the semantics and perspectives of individual datasets then a natural corollary follows; that sustainable data provisioning and aggregation must be a collaboration with data providers. The provisioning of data for aggregation systems cannot be top down and centralised but must include the expertise and knowledge of local experts who can ensure that data is represented correctly and to a high quality. The semantics of any individual dataset are unknown to an aggregator and this is the reason why fixed models have been employed – avoiding the need to engage in a more detailed way with data providers. The test of a data aggregation, with the objective of meaningful reuse, should be that a similar quality level (or higher) can be achieved when compared to a local provider service. A view of aggregation as a purely technical and mechanical assembly line, divorced from the providers will have no long term future. These conclusions have been clearly demonstrated by a whole range of projects over the last 20 years. (Oldman et al., 2014)

The Synergy Model

Providers are Primary

In many data integration projects the view and involvement of cultural organisations is one of content and service provider. This impression and role may be partly the result of institutional policies that have concentrated on narrow internal agendas rather than community wide concerns. However, as a result, cultural organisation are often treated as junior partners with no active part to play beyond supplying the raw materials. While it is true that cultural organisations are interested in wide dissemination of information many other factors contribute towards resource allocation decisions and organisations like museums, archives and libraries assess projects based on a wide range factors and potential benefits.

As the initial injection of project money invested in providers gradually ebbs away the longer term benefits of the project are required to 'kick in' and maintain interest for the provider. If the infrastructure treats the provider as a narrow and isolated service, and does not establish structures from which other benefits can flow, then the necessary energy needed from providers to sustain the venture will also ebb away. A data aggregation must be seen as a collaborative and community venture encompassing a wider range of outcomes. If outcomes are too divorced from providers then this represents a major flaw in the overall design. The Synergy model is designed to create an environment that can achieve a multiplicity of benefits because it is based around a mapping schema accessible to different kinds of expert (curators, technologists, data managers, etc.) on both the provider and aggregator sides, and because it is designed to treat the provider as both a contributor, collaborator and recipient of the service, rather than a mere supplier. Just as the cultural sector consists of many different organisations with different structures so the infrastructure of a community data aggregation initiative needs to allow for different requirements and the development of different communities. This is unlikely to be achievable with highly centralised projects divorced from local issues and with a narrow view of the role of providers.

Key Elements

The Synergy model describes three key aspects of data provisioning. These are:

1. The alignment of a provider's internal data model with an aggregator's target model.
2. The transfer of data to populate the aggregator's data repository.
3. The ongoing processes to maintain this alignment and the regular update of data.

Within these three key areas are some potentially difficult issues facing institutions particularly those with limited resources and technical expertise. Although the Synergy model defines all of the processes, roles and data objects necessary to support large scale quality aggregation, the underlying philosophy is to facilitate collaboration and the reuse of knowledge and expertise. In this respect the Synergy model is also the foundation for creating communities of people and tools and developing greater awareness of the potential of data aggregation.

An important aspect of Synergy is that while it provides a detailed treatment of the data provisioning process, it also ensures that data providers themselves are rewarded with significant tangible benefits. These include supporting outward facing digital strategies in research, education and engagement, but also leveraging the aggregation service itself to enrich internal data and improve internal processes. The idea is to stimulate an active provider community based around these benefits and empower providers to develop their own additional process extensions in addition to those specified by the Synergy framework. This includes overcoming issues faced by organisations

that have scare resources and limited budgets. In this way the Synergy model is not a 'one way' data provisioning service – it operates in many directions.

Joining a Synergy Based Aggregation

While Synergy is a reference model aimed at closer alignment with aggregation services there are practical issues to be faced in joining any data aggregation initiative. In many organisations the information systems (containing collection, archive, bibliographic and other cultural data) were never designed to be used in open integrated systems. In many cases they represent internal inventories or catalogues that require internal expertise to understand and use. When transferred to an open environment a range of data management issues can be exposed. This is true of the largest and the smallest cultural heritage organisation.

The issue of transferring meaning and context is addressed by the CIDOC CRM since mapping to it involves transferring not just data, but institutional knowledge about its underlying meaning and context. However, issues of data management often provide a far more practical problem. All cultural institutions have different types of data issues – data without proper validation, uncontrolled vocabularies, or fields that have been used for the wrong purposes. These are issues that are becoming more visible in any event as organisations publish data in different types of digital public forum. The activity of data provisioning, particularly through the Synergy model, will highlight these issues quickly and provide new and valuable information.

Synergy and Syntax Normalisation

The resolution of syntactic data issues is referred to in the Synergy reference as, "Syntax Normalisation". The data model or schema from a provider source system is visualised and associated data can be analysed to help identify potential problems. As issues arise new solutions and functions can be built to resolve these problems or to suggest different mapping approaches. The CIDOC CRM is designed to cover a range of generalisation and specialisation potentially offering temporary non-technical solutions. However, the intention is to create functionality around the Synergy process that allows syntax normalisation functions to stored and available for adaptation (syntax issues tend to vary between different institutions) and possible reuse for providers. This helps produce the best representation of the data possible. Ultimately, this is one area where organisations may have to resort to additional technical help, but also where the development of a connected community can be brought to bear. For example, the adaptation or configuration of tools could be achieved through real-time online support.

Synergy and X3ML – End to End Mapping

The Synergy system is different from many other aggregation systems because it encourages and supports a fuller representation of data than is generally undertaken. This is one reason why standard aggregation systems have limited scope. However, aligning a full source model to a target model that supports meaning and context can be less problematic than trying to decide how fields fit into artificial approximations within a smaller fixed field model (which has a lot in common with the received wisdom about 'round holes and square pegs'). Mapping, furthermore, only needs to be done once, and then can be used for a whole range of purposes or converted to any other cultural heritage standard. The Synergy model is based around a new language called 'X3ML', (Extract 3ples from XML). X3ML[2] is an XML based language which describes schema mappings in an accessible way so they can be created and maintained collaboratively by different types of expert including scholars

[2] https://github.com/delving/x3ml

and technologists, but still be machine readable. This more collaborative approach to defining data mapping instructions increases the quality and ensures that all relevant knowledge is used.

X3ML can provide instructions to convert data into many different formats such as W3C standards for the representation of the Resource Description Format (RDF), which are the basis of Linked Open Data. X3ML also provides a mechanism by which changes at either end of the aggregation relationships can be detected, reported on and corrected. The open and accessible nature of X3ML promotes the creation of different 'helper' components that may be developed by different suppliers or projects, but which can be used to help improve a Synergy based system. *However, the really important aspect of data provisioning that the Synergy / X3ML model seeks to address is to manage; the communication process for the mapping, the end to end provisioning pipeline, the error processing framework and finally the ongoing update of data.*

Synergy and Visualisation & Model Mapping

It is the ambition to encourage, through the Synergy/X3ML model, the development of non-technical visual tools for mapping. Traditional mapping or model alignment tools the user is confronted with terminology and functions that are unfamiliar exposing technical concepts that are difficult to grasp. Therefore efforts are underway to provide more graphical forms of mapping to visual 'patterns' that have been derived and categorised as a result of previous mapping work. The intention, facilitated by X3ML, is to produce a system that can quickly locate the correct pattern from a knowledge base or 'mapping memory' repository and use simple drag and drop techniques to align source schema with these target patterns. The patterns would have context sensitive help and additional functions for adding optional elements to a core pattern.

Synergy and URI Schema

The Synergy model is aimed at aggregations that retain the identity of the provider organisation in the URI (Uniform Resource Identifier). For example the URI for the British Museum's Easter Island Statue, Hoa Hakananai is,

http://collection.britishmuseum.org/id/object/EOC3130.

The acquisition event has the URI

http://collection.britishmuseum.org/id/object/EOC3130/acquisition.

The URI scheme was determined by the British Museum but many aggregations will have their own URI policy specific to their aggregation system. By only specifying a URI policy at the level above the domain name of the provider it is clear from which organisation the data originated. Using a consistent set of aggregator URIs above the domain name means that aggregators do not have to co-reference across different URI implementations. The Synergy system will allow the application of URI policies as specified by a particular aggregator and also facilitate co-reference across the integrated datasets. The implementation of URIs creates another process point in the Synergy model to identify data management issues and additionally check and verify external URIs.

Terminology

The CIDOC CRM employs meaning through relationships and types. Cultural heritage organisations qualify certain fields and concepts by using terminologies. These terminologies can affect the way that the provider's model maps to the aggregator's model. For example, it can affect the visual

pattern (described earlier) that is used for a particular mapping. The use of the term, “donated by” would require a different mapping pattern to the term, “donated through” because one is a direct acquisition (directly *from* an individual or group) and one is indirect (*through* an agent).

Aggregators are likely to have a consistent set of terminologies integrated into their system. These will be different from those used by the different data providers. The Synergy system defines a process of co-reference, in most cases to the aggregator’s broader terms, and where terms don’t exist a creation event would be initiated. To support this area of the Synergy process model algorithms are being developed that make use of both CIDOC CRM semantics as well as string matching techniques to determine most likely fit, and present these to users with the associated CIDOC CRM based evidence.

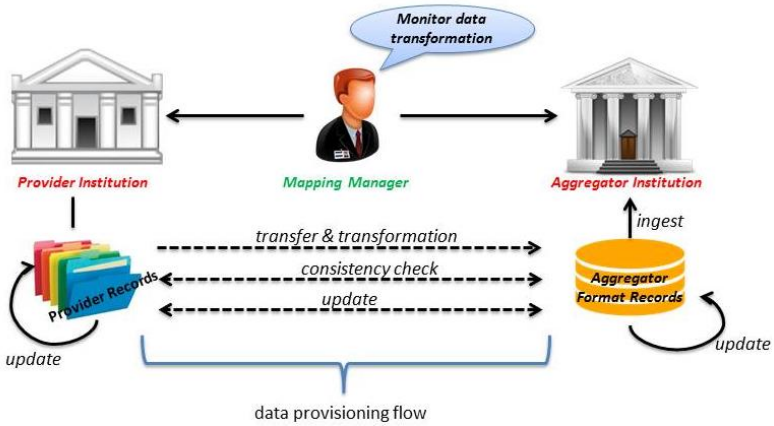
Transfer & Transformation

Once models have been aligned and all data errors have been resolved the data can be transferred to the aggregator. Synergy recommends that the raw data also be transferred so that information can be recovered without any need for action by the provider. The model includes processes for validating identifiers and creating modification dates. The transferred data is then transformed into the aggregator’s model using the X3ML instructions. This again may identify further errors that might be resolvable by the aggregator, or might need to be referred back to the provider.

An aggregator may then use algorithms that attempt to find equivalences for people and places so that these URIs can be mapped or normalised. The question of whether versions of data are kept, as new data enters the system, depends upon the services that the aggregator provides. From this point onwards the updates are provided according to a terms agreed by the parties. If changes are made, for example, to the provider’s or aggregator’s models, then the relevant processes are automatically revisited (changes can be automatically reported) and the provisioning re-aligned in terms of both models, data, co-referencing and instance matching.

Synergy Roles and Responsibilities

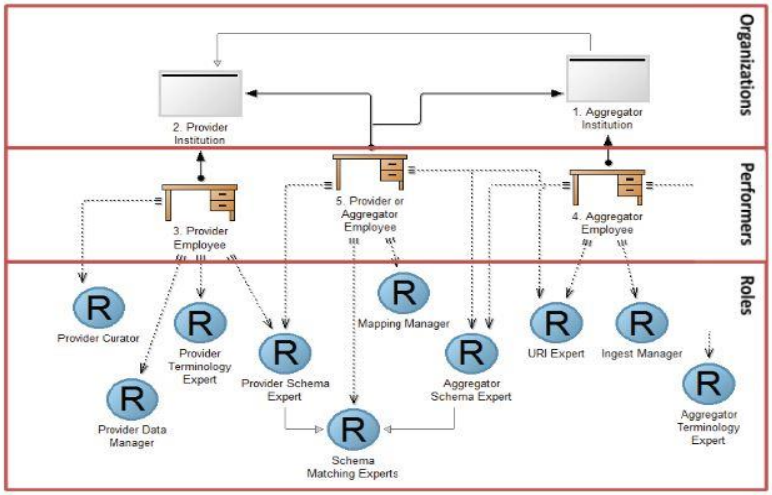
In the real world cultural heritage datasets are managed within a diverse range of structures. Local museums may rely on the technical services of local authorities, or may have their own in-house technology support. Curators may participate in documentation roles or it may be fully resourced by documentation specialists. Suppliers may be contracted to provide additional consultancy services either on a regular or on a time-and-materials basis. The Synergy model defines a set of roles that may be distributed and undertaken in different ways depending on these different structures. Multiple roles might also be fulfilled by the same people or groups. In some cases, although a role might be expected to be a provider role, it may in practice be pragmatically satisfied by the aggregator (if it involves knowledge available to the aggregator). Apart from the two main parties themselves (the Provider and the Aggregator), the key role in managing the relationship is the ‘Mapping Manager’.



In the Synergy model the relationship between provider and aggregator extends beyond a simple agreement to participate, and a permission to use the data. The management of the relationship in Synergy extends to a more detailed plan of how the data provisioning is achieved and operated and a set of terms that are specific to a particular institution and their needs - as well as the needs of the aggregator.

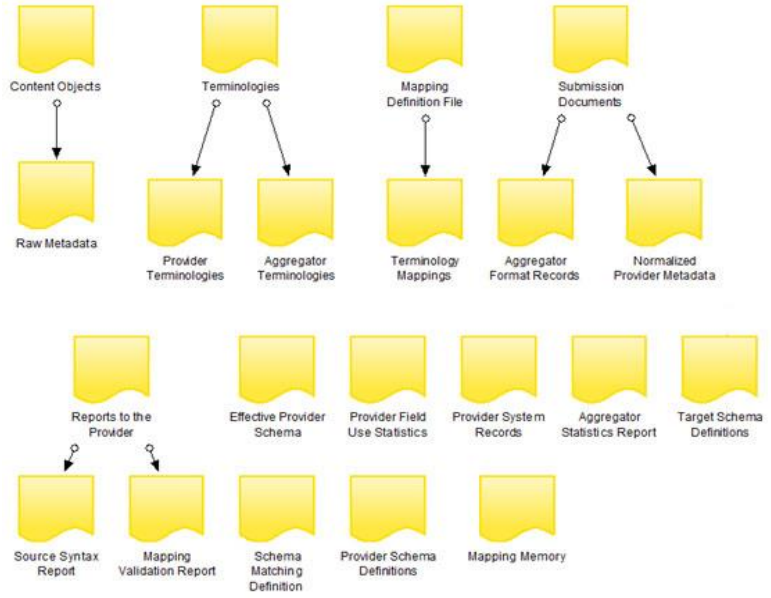
Under this management framework other expert roles include:

- **Provider Data Manager** – dealing with technical and system issues;
- **Provider Schema Expert** – able to provide information on how fields, tables and elements are used in practice locally;
- **Provider Terminology Expert** – an expert in local terminologies;
- **Aggregator Schema Expert** – able to provide information on the target model (for example the CIDOC CRM);
- **Aggregator Terminology Expert** – expert on the set of target terminologies used as a reference for local terminologies;
- **Ingest Manager** – manages the transfer of data and population of the aggregators integrated data repository;
- **URI Expert -** needed to produce and maintain a URI policy for the data;
- **Schema Matching Experts** – experts in the respective schemas necessary to align provider and aggregator models.



These roles are separated out for the purposes of defining the Synergy model, but in reality can be performed with relatively small levels of resource because of the availability of knowledge resources (developed by existing CIDOC CRM implementers) that will be available to a Synergy based aggregation.

Synergy Data Objects



The Synergy reference model relies on a number of inputs and outputs. Importantly, Synergy is optimised towards feeding information back to providers not just in terms of errors and data issues, but also important statistics and analysis. Components that are developed around this information will be encouraged to visualise it further to the extent that a Synergy service is likely to inform organisations about their data in ways not available from their internal information systems. Data objects include the information submitted to the Synergy system by the provider, like structured data, unstructured data (documents and images), and terminology, as well as the schema being used for the mapping. What Synergy produces can be used by organisations to improve their data, whether for the data provision service, or for other systems and internal improvements. Reports and statistics cover mapping and syntax validation as well as ingest and co-referencing. Crucially it also provides information about the mapping definitions themselves which can be analysed and used to improve the "mapping memory" service. As a result of using the system organisations will collect invaluable information about the condition and quality of their data that can be used for a wide range of purposes.

Information System Objects & CultureBroker

In order to implement the Synergy process model efficiently software components are required. It is not expected that there will be a single set of components that everybody uses. Instead different projects and aggregations will develop tools aimed at addressing the Synergy model in different ways, to suit different requirements. However, these tools would be available for other projects to select. For example, different visualisation and mapping tools with particular specialisations may all use the X3ML standard and be compatible with Synergy.

Some of these components are already being developed by the CultureBroker[3] project, a collaborative initiative based in Sweden involving the Swedish Arts Council, Swedish National Archives, ResearchSpace (British Museum)[4], Delving BV[5] and F.O.R.T.H. (Institute of Computer

[3] http://culturebroker.eu

[4] http://researchspace.org

Science)[6]. The project is currently creating a data aggregation for a number of regional cultural organisations in Sweden mapping both collection and archive data to the CIDOC CRM standard and it is intended that this system will scale up to larger aggregation with a number of different services including CultureCloud[7] and ResearchSpace components.

The anticipated Synergy toolset would include components for the following functions;

- Data Analysis
- Syntax Normalisation
- Terminology Mapping
- Schema Matching
- Link Checking
- X3ML Transformation
- Instance Matching

Many of these tools, following the Synergy model, would provide reports and information that can be accessed by both the Provider and the Aggregator.

Future Work & How to Help

If you are interested in learning more about the Synergy model or contributing towards shaping the specification then the following links are relevant.

- A working draft of the Synergy Model now resides on the CIDOC CRM site at;

  http://www.cidoc-crm.org/working_editions_cidoc.html

- A direct link to the document is;

  http://www.cidoc-crm.org/docs/SRM_v0.1.pdf

- Realizing Lessons of the Last 20 Years: A Manifesto for Data Provisioning & Aggregation Services for the Digital Humanities (A Position Paper)
  http://www.dlib.org/dlib/july14/oldman/07oldman.html

- Comments on the paper can be emailed to;
  crm-sig@ics.forth.gr

Events and workshops about the Synergy model will be available at;

- http://www.cidoc-crm.org/special_interest_meetings.html

[5] http://delving.eu

[6] http://ics.forth.gr/

[7] http://culturecloud.eu/

References

[1] Bearman, D. (2010) 'Standards for Networked Cultural Heritage', in Ross Parry (ed.) *Museums in a Digital Age*. London; New York: Routledge.

[2] Crofts, N. et al. (eds.) (2011) *Definition of the CIDOC Conceptual Reference Model - cidoc_crm_version_5.0.4.pdf*. [online]. Available from: http://www.cidoc-crm.org/docs/cidoc_crm_version_5.0.4.pdf (Accessed 14 May 2014).

[3] Doerr, M. (2003) The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI magazine*. 24 (3), 75.

[4] Doerr, M. & Crofts, N. (1998) *Electronic espernato—The Role of the oo CIDOC Reference Model*. Citeseer. [online]. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.47.9674&rep=rep1&type=pdf (Accessed 26 August 2013).

[5] Janes, R. R. (2009) *Museums in a troubled world: renewal, irrelevance or collapse?* London; New York: Routledge.

[6] Oldman, D. et al. (2014) Realizing Lessons of the Last 20 Years: A Manifesto for Data Provisioning & Aggregation Services for the Digital Humanities (A Position Paper). *D-Lib Magazine*. [Online] 20 (7/8), [online]. Available from: http://www.dlib.org/dlib/july14/oldman/07oldman.html (Accessed 15 July 2014).