Europeana – Core Service Platform

MILESTONE

MS1:Specifications for the accurate representation of data providers' names in the DSI

Revision	1
Date of submission	24 of November 2015
Author(s)	Pablo Uceda Gómez (EF)
Dissemination Level	Public

Co-financed by the European Union
Connecting Europe Facility

REVISION HISTORY AND STATEMENT OF ORIGINALITY

Revision History

Revision No.	Date	Author	Description
0.1	18/11/15	Pablo Uceda Gómez (EF)	Initial draft
0.2	20/11/15	Pablo Uceda Gómez (EF), Valentine Charles (EF), Henning Scholz (EF)	Second draft, comments integrated
1	22/11/15	Pablo Uceda Gómez (EF),Henning Scholz (EF)	Final comments integrated, layout edits, minor changes

Statement of originality:

This milestone contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

The sole responsibility of this publication lies with the author. The European Union is not responsible for any use that may be made of the information contained therein.

Europeana DSI is co-financed by the European Union's Connecting Europe Facility’

Table of Contents

1. The Europeana Aggregator Model and the current representation of the data chain in EDM.

Europeana is the network for the cultural heritage sector in Europe, and shares a vision of the world where every citizen has access to all cultural heritage. Europeana currently gathers more than 47 million records form thousands of European cultural an scientific institutions.

With the available resources it is not possible for the Europeana Office to work directly with all the organisations that are providing data to Europeana. The Europeana ecosystem relies on a network of national, thematic and domain aggregators instead. An aggregator is an organisation that collects, normalises and enriches data from multiple institutions.

The aggregator model has made it possible to obtain metadata from thousands of cultural heritage and scientific organisations while directly receiving data from less than 150 direct providers.The Europeana network includes three main types of aggregators:

● **National aggregators**, that gather data from a specific country or region and whose data providers are situated within that geographical area.

● **Domain aggregators**, that collect data from a particular sector (such as museums, archives or libraries) and whose data providers may be international.

●**Thematic aggregators**, that are focused on a particular topic or theme (such as fashion or food and drink) and whose contributors may be located in more than one country.

In the basic aggregator model, data is flowing from the source institution via an aggregator to Europeana. The aggregator cleans, and enriches data before submitting it to Europeana. Finally after data has been processed by the Europeana Data Partner Services Team, the data becomes available to the user via the Europeana Collections site and the Europeana API.

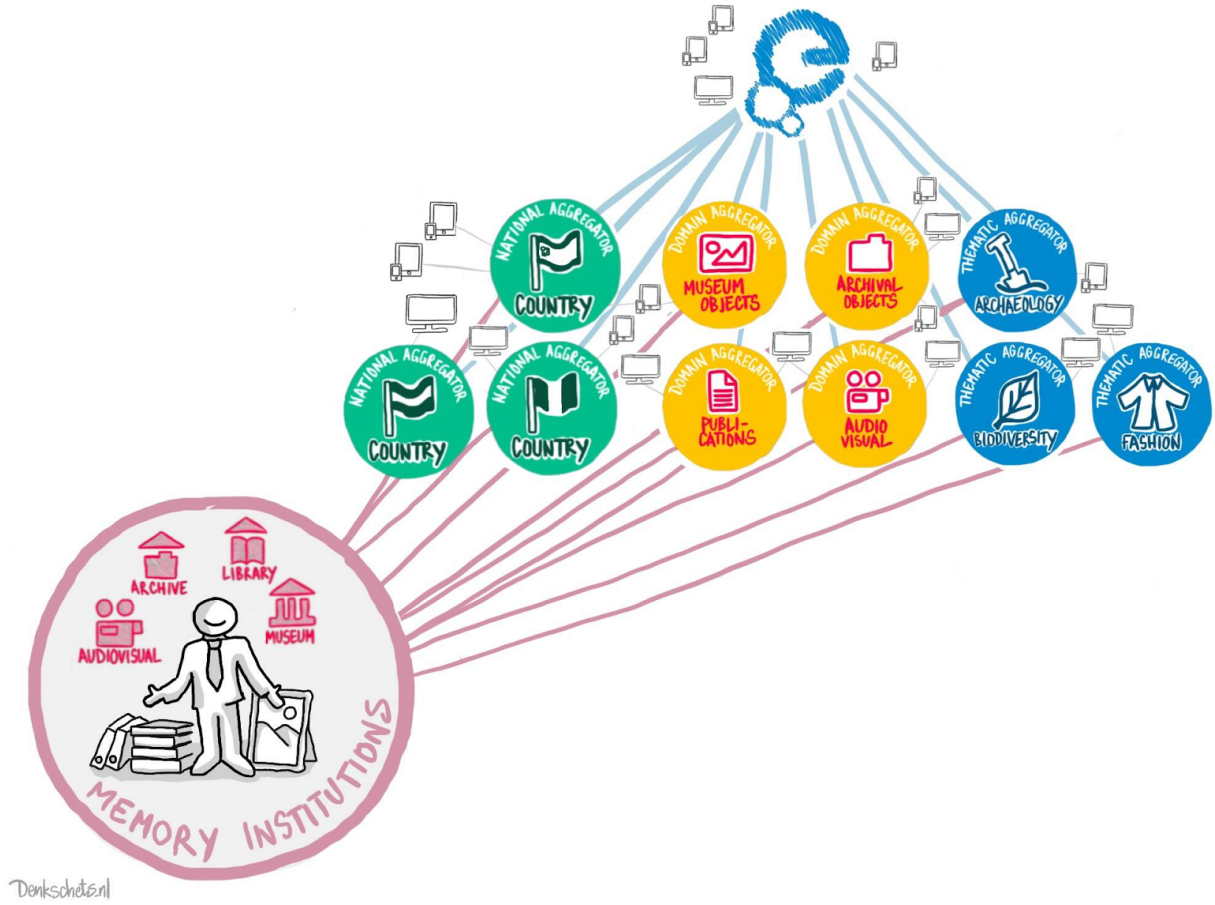

Fig. 1. Current (but simplified) aggregator model of Europeana: national aggregators (green), domain aggregators (yellow), thematic aggregators (blue). It is simplified in the sense that e.g. sub-aggregators are not included in this representation of the model.

The Europeana Data Model (EDM) represents the basic aggregation chain as shown in figure 1, using two repeatable and non mandatory properties.

1. **edm:provider**: This property is used to describe the entity that provides the data directly to Europeana. In most of the cases it is a national or domain aggregator like Hispana or a project like EAGLE.

2. **edm:dataProvider**: This property is used to describe the institution that has provided, and usually generated, the data, in most of the cases a museum, library or archive. For example the Rijksmuseum.

After the ingestion in Europeana, both the edm:provider and the edm:dataProvider generate a facet in the left-hand side of the Europeana user interface that enables the user to select the records form a particular institution or project(Fig 2).

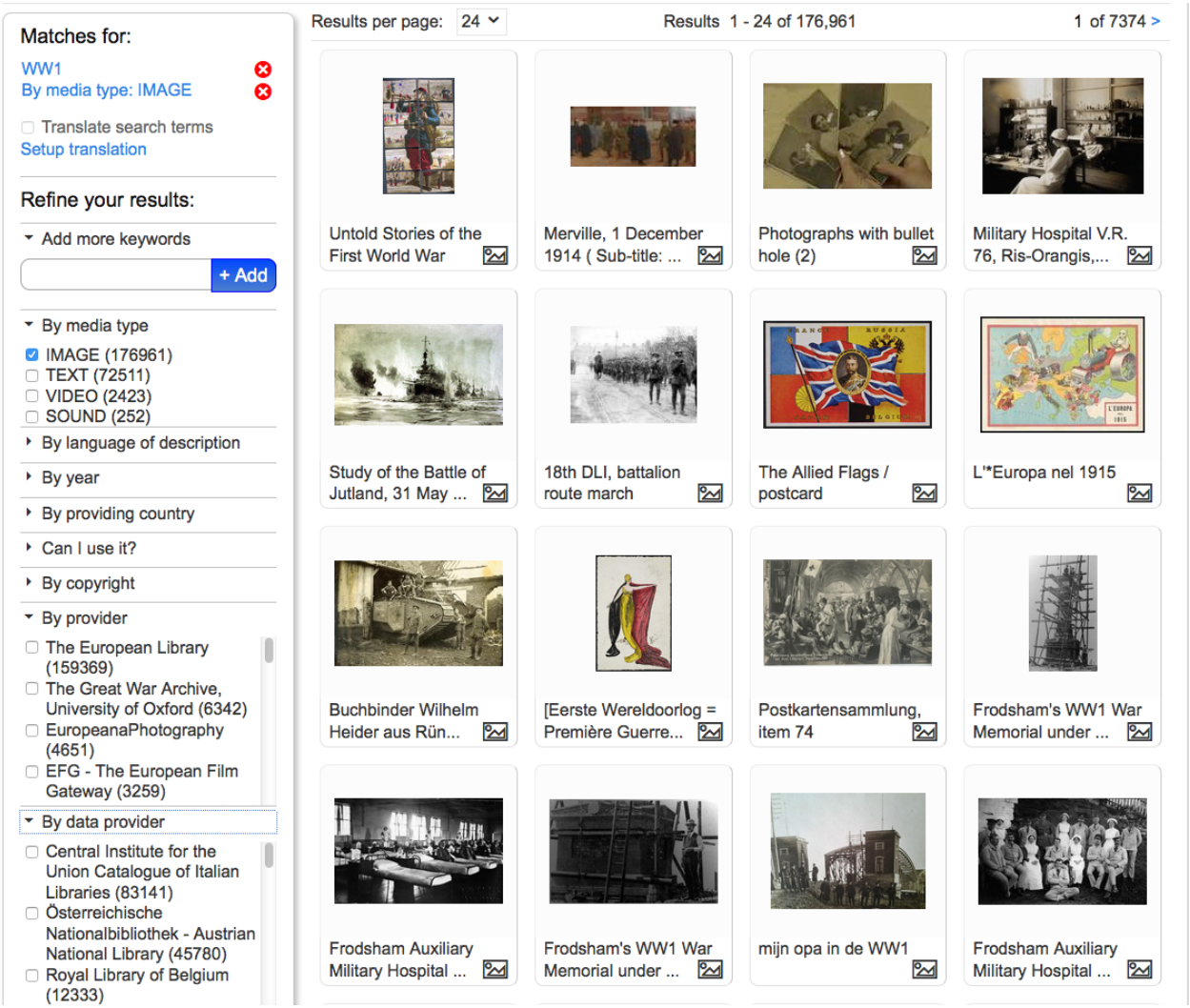

Fig. 2. Screenshot of europeana.eu with the two facets for edm:provider and edm:dataProvider being expanded.

In spite of this theoretical simplicity, the data chain is often more complicated and frequently there are several layers of aggregation in the data flow[1].

As a consequence of the evolution of the aggregation landscape and the increasing complexity of the data chain, Europeana is not able to represent accurately the aggregator ecosystem and the relationships between institutions, aggregators and projects.

This inadequate representation of the data providers together with lack of a system to avoid mistakes or inconsistencies in the values that populate the edm:dataProvider field, makes it almost impossible even for Europeana to know exactly the total number of institutions that are delivering data to us.

[1]More information in "Recommendations to Improve the Aggregation Infrastructure"

This report identifies the problems that are blocking Europeana from accurately representing data providers’ names in the DSI and proposes short and mid-terms solutions in order to start designing a better representation of the data chain and therefore increase the visibility of all the data providers.

2. Problems in the identification of Europeana data providers

In June 2015 the edm:dataProvider field was populated with 3,816 different values. Approximately 16% of these values are repetitions, inconsistencies or human mistakes. An unknown number of data providers is hidden under one or more layers of aggregation. At this stage it is difficult to give a precise figure of the number of data providers but the causes of this inaccuracy and ways to solve them will be identified herein.

2.1 Lack of normalization

The edm:dataProvider field is a free text property. That means that the data provider can populate this field with whatever value they might like to use. The lack of normalization leads to the appearance of a number of imprecisions and inconsistencies in the data that are described in the following sections.

2.1.1 Structure of organisation names

In some cases when delivering data, the provider populates the edm:dataProvider field with different values for the same providing institution.

Number of records	Value
3,243	Bodleian Libraries, Oxford University
354,441	Bodleian Libraries, University of Oxford

2.1.2 Units of organisation

Ideally the edm:dataProvider value should be representing an entity with legal personality that has signed the Data Exchange Agreement although there is no hard rule for this.

Sometimes data providers fill this value with sub-sections or departments of their own infrastructure that may or may not be separate entities.

Number of records	Value
1,850	Victoria and Albert Museum
3	Theatre and Performance Department, Victoria and Albert Museum
5	Sculpture Department, Victoria and Albert Museum

1	Metalwork Collection Victoria and Albert Museum
5	Indian Department, Victoria and Albert Museum
3	Far Eastern Department, Victoria and Albert Museum
4	Ceramics Department, Victoria and Albert Museum

2.1.3 Multilinguality of organisation names

Due to the international nature of Europeana, it is common that the data provider expresses its name in its native language or in English, or both. The edm:dataProvider is a unique value so it cannot be provided twice with different language tags.

Number of records	Value
725	Max Planck Institute for the History of Science
1,288	Max-Planck-Institut für Wissenschaftsgeschichte
1,334	Max-Planck-Institut für Wissenschaftsgeschichte (Max Planck Institute for the History of Science)

The use of various alphabets could be another source for the diversity of names.

Number of records	Value
22,217	Регионална библиотека ПЕНЧО СЛАВЕЙКОВ - Варна
12,264	Регионална библиотека ПЕНЧО СЛАВЕЙКОВ - Варна / Public Library - Varna
10,285	Регионална библиотека ПЕНЧО СЛАВЕЙКОВ - Варна/Public Library - Varna

2.1.4 Use of acronyms

Frequently providers use an acronym to represent the name of its institution. This alone or in combination with other aspects can increase the diversity of names being present.

Number of records	Value
14,716	CIMEC
857	CIMEC - Institute for Cultural Memory
11,401	cIMeC - Institutul de Memorie Culturală

2.1.5 Use of links as data provider name

Sometimes providers give the link to the web or on-line portal of the institution instead of using their real name.

Number of records	Value
121	eSbirky
44,670	eSbírky
1,388	www.esbirky.cz

In other cases, links are resolving to web portals. It becomes difficult to specify if they are institutions themselves or if they are just showcasing the content of other organisations.

Number of records	Value
148,375	askaboutireland.ie
227	www.maltamigration.com

2.1.6 Human mistake

Unintentionally, the field could just be filled with the wrong values. The Operations Officers do its best to avoid ingesting wrong data but sometimes the mistakes go through undetected.

Number of records	Value
768	General Secretariat for Culture, Hellenic Ministry of Education and Religious Affiars, Culture and Sports
161	General Secretariat for Culture, Hellenic Ministry of Education and Religious Affiars, Culture and Sports Γενική Γραμματεία Πολιτισμού, Υπουργείο Παιδείας και Θρησκευμάτων, Πολιτισμού και Αθλητισμού

2.2 Intermediate providers

The data is not always provided directly from the data provider to an aggregator that is providing the data to Europeana. It is common that other organisations act as intermediaries.

For example, in Spain the data from the museum domain is collected by CER.ES, an organisation representing 86 different institutions, and then sent to Europeana via Hispana, the Spanish national aggregator. Currently the institutions that are providing data to CER.ES are not represented in the edm:dataProvider field (see Fig. 3).

Up to now, to solve this situation the name of the data provider has been included in other fields such as **dc:source**, **dc:rights**, **dc:contributo**r , **dc:terms:IsPartOf** or **dc:publisher**.

This problem is stopping us to be able to retrieve automatically all the data providers values and to represent all the data providers in the facets of the user interface.

Fig. 3. The data from this record has been provided by el Museo del Traje (Madrid). In this case this info is represented in dcterms:IsPartOf.

In other cases the provider chooses to specify both the actual dataProvider and the intermediate provider in the edm:dataProvider field. In that way neither the intermediate provider nor the data provider have a facet of their own.

Number of records	Value
1,610	The Cyprus Institute - STARC
870	The Cyprus Institute - STARC / Art Gallery of Archbishop Makarios III Foundation
53	The Cyprus Institute - STARC / Byzantine Museum of the Archbishop Makarios III Foundation
203	The Cyprus Institute - STARC / Mediterranean Archaeological Research Institute-Vrije Universiteit Brussel
1	The Cyprus Institute - STARC / Βυζαντινό Μουσείο Ιδρύματος Αρχιεπισκόπου Μακαρίου Γ
732	The Cyprus Institute - STARC / Βυζαντινό Μουσείο Ιδρύματος Αρχιεπισκόπου Μακαρίου Γ΄
207	The Cyprus Institute - STARC / Πινακοθήκη Ιδρύματος Αρχιεπισκόπου Μακαρίου Γ΄
130	The Cyprus Institute - STARC, Cyprus Folk Art Museum
7	The Cyprus Institute - STARC/ Department of Antiquities, Cyprus
1	The Cyprus Institute - STARC</source>
2	The Cyprus Institute - STARCThe Cyprus Institute - STARC
127	The Cyprus Institute-STARC, Cyprus Folk Art Museum

3. Outlining possible solutions

All the problems described above make it difficult for the end-user to identify which organisation is providing data to Europeana. They also diminish the visibility of the data provider and make it impossible for Europeana to have a precise count of how many institutions are delivering data. The problem is even more complicated due to the dynamic and evolving nature of the data flow and the constant changes in the database. Due to the diversity of the problems, different approaches are necessary to address them. They are explained in the following, together with a timeframe for implementation.

3.1 Best practices

The first step to tackle inconsistencies in the data provider names will be to establish a set of recommendations and guidelines about how to best populate the edm:dataProvider property, i.e how to best deal with multilinguality, acronyms, how to avoid ambiguities etc. This will require a communication and coordination effort with data partners. The documentation will be updated accordingly and new data partners will be advised to follow them.

Besides improving the consistency of new data the improvement of the legacy data in the Europeana database needs to be addressed. It could be updated manually but this is a very time-consuming process that would overload the Europeana Data Partner Services team. Instead, we suggest to start a gradual update process that gives the initiative to the data providers themselves. In that context the new Europeana Statistics Dashboard will play a key role.

Europeana is preparing a range of improvements that will affect the technical infrastructure, the user interface and the ingestion process itself. Among them is the new **Europeana Statistics Dashboard** that provides key metrics and visualizations about the data sets such as the traffic, information about data providers and content.

With the use of this new tool it will be very easy for data providers to detect mistakes and inconsistencies in the data provider values. We will encourage our partners to report any issue to request an update of their data. The Data Partner Services Team will liaise with data partners to implement the desired changes.

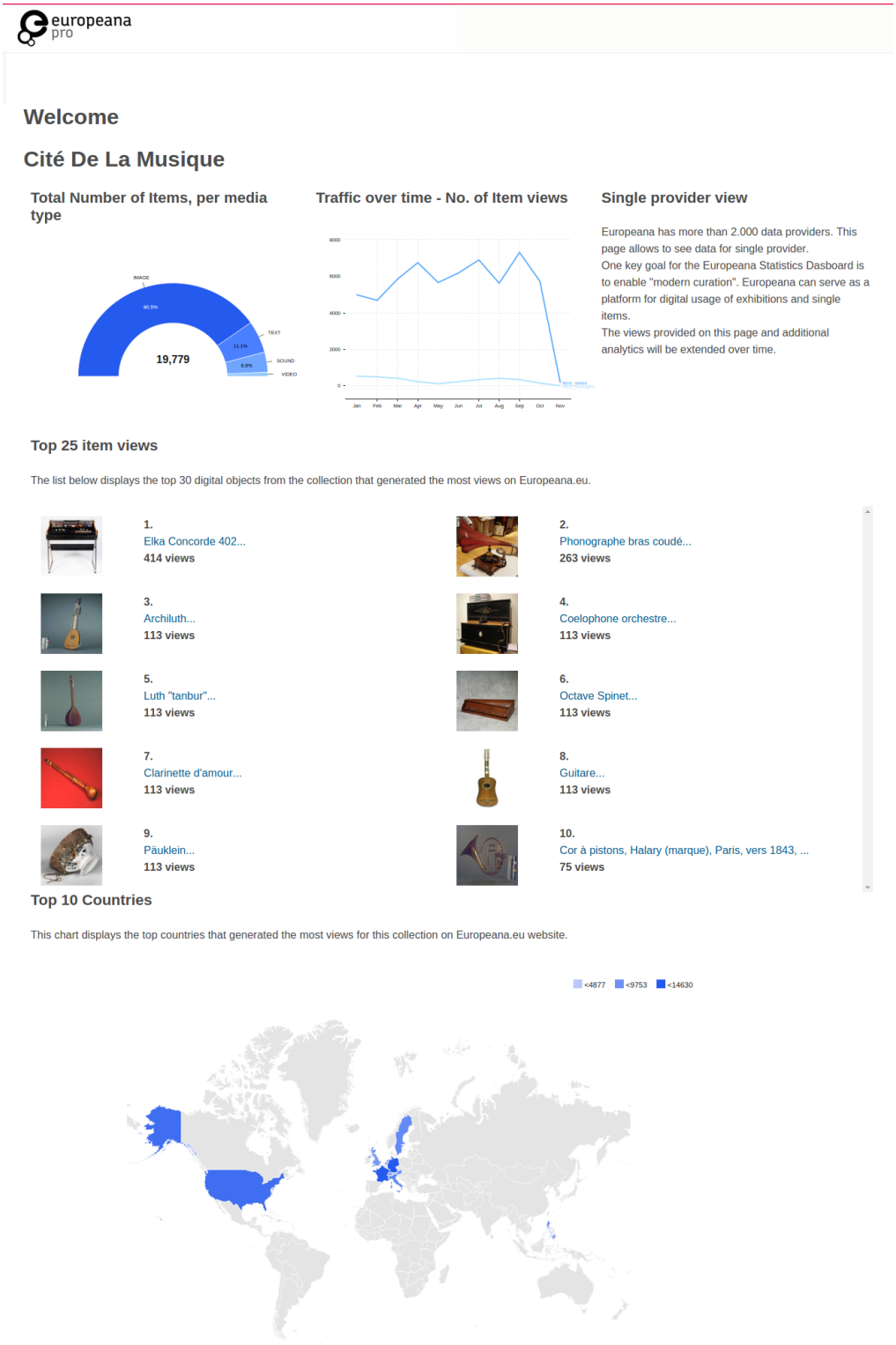

Fig **4.** A screenshot of the test version of the dashboard.

3.2 EDM Profile for Organisations

All the solutions described above could improve the representation of data providers and the quality of the data. Anyhow they cannot solve in a definitive manner the appearance of inconsistencies, mistakes or ambiguities.

In order to solve the lack of normalization of the edm:dataProvider field and to get the information about provider and data providers in a controlled fashion, in 2013 the Europeana R&D team started to work on the new edm Organisation Profile for providers and data providers.

Currently providers and data providers are represented as a text string. We would like to represent them as real entities. Representing the providers and data providers that way will allow us not only to achieve normalization but to use controlled vocabularies and get more controlled information about the providing institutions. The EDM profile for organisations will work in a similar way as other contextual resources already in use such as edm:Agent[2].

This profile has not been yet implemented but it summarises the direction that Europeana wants to follow.

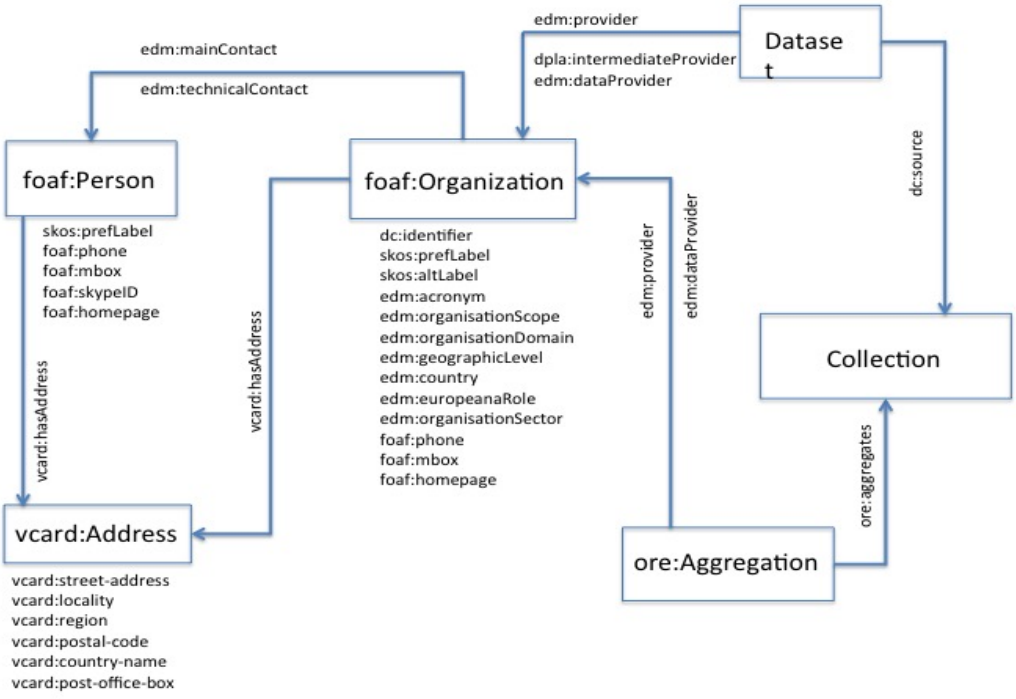

Fig **5**. This chart summarises the properties in the new edm profile for organisations.

[2]To see more information about how entities work in EDM please see the "Europeana Mapping Guidelines"http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Mapping_Guidelines_v2.2.pdf

In order to normalize the provider names in Europeana, the profile for organisations could be used in two different ways:

- To perform an alignment with the data in the Europeana Customer Relationship Management system (Sugar CRM).
- To fetch data from a controlled vocabulary for organisations

3.2.1 Alignment with the Europeana Customer Relationship Management system (Sugar CRM).

SugarCRM is the system that Europeana uses to register information about delivering organisations and their datasets.

In the CRM, Name, Alternate name, Acronym, Role/Relationship to Europeana, Thematic Partner Network, Europeana Network, API status are mandatory properties. Except the last three, these properties have its equivalent in the fields of the EDM profile for organisations.

Once the profile has been implemented, we would be able to fetch the data from the CRM and align it with properties of the profile. This process will require a clean-up of the data of the CRM and to establish coordination between R&D, Data Partner Services Team and the development team but it will allow to standardize the data provider values of the whole database at once.

3.2.2 ISNI

ISNIAnother possibility to normalize the name of the organisations in Europeana is the use of controlled vocabularies. The new profile will allow to include a URI from a controlled vocabulary as the identifier of the foaf:Organization class and fetch the information from the target vocabulary. Currently Europeana R&D is considering the use of ISNI (International Standard Name Identifier).

 is aiming to create an international authority file that gives to the represented entities a unique persistent identifier. ISNI gathers data from hundred of databases worldwide[3]. ISNI Is currently hosting public records of 8.99 million entities, including 525,636 organisations. ISNI also aggregates information about the organisations such as the Prefered Label, different variants of the name, such as acronyms, translations etc. With use of ISNI identifier we would be able to incorporate all of this information in the data.

[3] To know more about how ISNI works you can read http://isni.org/how-isni-works y

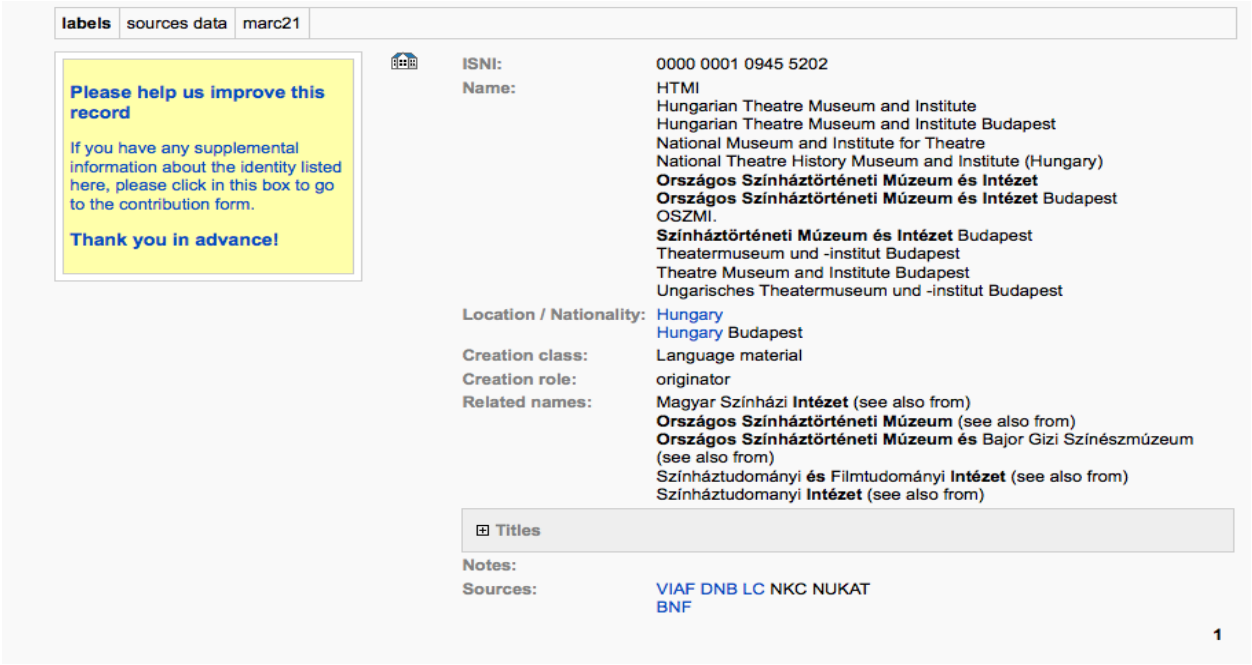

Fig 6. ISNI register for The Hungarian Theatre Museum, one of European Data Providers. In the record there is additional information such as translations or alternative names that we will be able to fetch.

Before stating to use ISNI, it is necessary to test the degree of representation and accuracy of SINI in relationship with the institutions that are delivering data to Europeana.

We have performed a first manual check in the ISNI database with a sample of 100 data providers. The result of this test shows that 56% of the sample have an ISNI ID. We have found out some inconsistencies and ambiguities in the records, e.g. institutions having more than one entry in ISNI. Further research will be performed to know the suitability of ISNI for Europeana.

It is important to say that ISNI is an evolving system. Multiple partners are cooperating with an ISNI agency to improve the quality and accuracy of the data so it is expected that with the pass of time the amount of institutions and the quality of the data will be improved.

In order to improve the data and to incorporate new organisations, Europeana might consider in the future to become an ISNI registration agency. This will allow Europeana to register in ISNI those data providers that are not yet registered.

C.3 Extending EDM with dpla:intermediateProvider

As previously described, the Europeana data chain is much more complex than the simple aggregator-data provider relationship. In order to improve the representation of the data flow, the Europeana R&D team and our development team will extend EDM with a new property, the **dpla:intermediateProvider**[4]. It will be non mandatory and repeatable and it will be ready to use at the end of the first year of Europeana DSI.

The institutions that are currently hidden under sub-aggregators or other intermediate providers will be the main beneficiaries of the use of dpla:intermediateProvider. Once implemented, the first step will be to identify which data providers are acting as an umbrella for multiple organisations. We will examine our database and shortlist those organisations that could benefit from its use. Secondly, we will approach them to examine how complex it is to identify all the data providers within a sub-aggregator or other organisations. The last step will be to reprocess the updated data and steadily start to represent the hidden organisations.

[4]This property is re-used from Digital Public Library of America EDM profile http://dp.la/info/wp-content/uploads/2015/03/MAPv4.pdf .

4. Conclusions

- There is an urgent need to identify precisely all the data providers that are delivering to Europeana and to improve the representation of the data chain in edm.

- The normalization of the data provider names and its proper representation is a complex problem with no immediate solution. In the short term, cooperation and communication with data partners is the best way to detect and steadily correct inconsistencies and ambiguities. The new Europeana Statistics Dashboard will play a key role in this process.

- In a longer term the definitive normalization of the data provider values will require the implementation of the EDM profile for organisations. Once that happened there are two ways to proceed.1) To perform an alignment between Europeana CRM and the data in the portal, 2) To fetch data from controlled vocabularies. In this sense Europeana have started to test the use of ISNI.

- The representation of data providers that are currently hidden under other umbrella organisations will be solved with the implementation of the new dpla:intermediateProvider field. This update will require communication and coordination efforts with partners to identify all the institution that are currently not well represented, modify the data and re-ingest the sets.