Using Dublin Core Standard for the Metadata Description of Transport Statistics—Practical Experience from a Project Dedicated to the Set-Up of an Interlinked Statistics Portal

Rene Kelpin
German Aerospace Center,
Germany
Rene.Kelpin@dlr.de

Antje von Schmidt
German Aerospace Center,
Germany
Antje.vonSchmidt@dlr.de

Michael Hepting
German Aerospace Center,
Germany
Michael.Hepting@dlr.de

Petra Kokus
German Aerospace Center,
Germany
Petra.Kokus@dlr.de

Alexandra Leipold
German Aerospace Center,
Germany
Alexandra.Leipold@dlr.de

Thorsten Schaefer
German Aerospace Center,
Germany
Thorsten.Schaefer@dlr.de

Abstract

The analysis of data on transport developments is one major objective in the transport research focus of the German Aerospace Center (DLR). Following this objective, the DLR's Institute for Transport Research (VF) offers, together with the Clearing House of Transport and Mobility, a unique collection of publicly funded travel and mobility surveys for Germany and is the official provider for German household surveys and statistics. Additionally, information about similar statistical data sets and data portals in Europe is made available. With the "MONITOR" portal, the Institute of Air Transport and Airport Research (FW), as the second DLR institute in the aforementioned field, provides detailed information and statistics concerning air transport developments. Furthermore, with a set of indicators the institute performs analyses of the global long-term air transport development regarding air traffic and financial performance besides sustainability issues. As both institutes use the Dublin Core Standard for the description of the data sources in use, in 2011 the idea came up of realising a common (meta-) data repository for interested users who have the need to combine and investigate different transport statistics. Accordingly, the project "STRADA@DLR" (Search TRAnsport DAta @ DLR) was launched in cooperation with DLR's Facility on Simulations and Software Technology (SC) to create an external search and analysis system allowing directed access to the mentioned data repositories. In this context, the presented project report discusses the usage of Dublin Core Standard in both institutes, as well as the organisational challenges and the technical approach in order to elaborate a harmonised metadata scheme for the implementation of the STRADA@DLR portal.

Keywords: transport statistics; metadata quality; metadata interoperability; integration of metadata and ontologies.

1. Introduction and Motivation

Besides research in aeronautics, space, energy and security, transport research is one core element of the German Aerospace Center research portfolio. Within this field, the Institute for Transport Research (VF) and the Institute of Air Transport and Airport Research (FW) are dealing with data management and knowledge transfer, which is a crucial prerequisite to support the elaboration and dissemination of transport research results. Addressing the need for adequate data identification, storage, descriptions and (re)usage, both institutes have built separate data repositories over the past years which are also available for external users.

The data collections concerned are both set up on the Dublin Core Standard as the basis for the underlying metadata scheme to classify and characterise the data sources which are stored in the separate data repositories. However, due to different needs of the separate institutes (see following sections) there was the need for a partial deviation from the Dublin Core Standard. In 2011 the project STRADA@DLR was initiated following the objective of bundling the data management work and expertise of VF and FW with their separate data collection activities at the Clearing House of Transport and Mobility (http://www.dlr.de/cs) and the MONITOR portal (http://www.airtransport-monitor.eu).

A special benefit for transport researchers should be generated through allowing this target group direct access to data descriptions concerning various transport statistics and studies in the field of air transport and personal mobility affecting mainly road and public transport. Intelligent search facilities based on a common and harmonised metadata concept in the background should, furthermore, facilitate the process of identifying the appropriate data, assessing their value with regard to specific research questions and recognising new options for combining different sources in order to improve the own research results.

Given this task, the following two sections describe in a first step the Dublin Core Standard with regard to its different usage in both mentioned institutes. In a second step, it is discussed in sections 4 and 5, how the different metadata concepts of VF and FW were merged from an organisational and technical point of view. The conclusions that can be drawn from this process can serve as useful examples for similar activities in the field of metadata interoperability. Lessons learned will be derived from faced logical difficulties and technical barriers.

2. Metadata Concept of the Clearing House of Transport and Mobility

Within the Institute for Transport Research, the Clearing House of Transport and Mobility (Clearing House) acts as a collector and provider of transport and mobility studies and statistics. Amongst other material, the Clearing House holds and provides raw data statistics such as the "German Mobility Panel" (1994 - 2012) and the largest German household survey, "Mobility in Germany" (2002 & 2008). One additional focus of the Clearing House deployment is placed on the cooperation and linkage with existing data repositories and portals dealing with further aspects of transport research.

For the description of metadata elements of statistics and studies provided, the Clearing House applies a format based on Dublin Core (DC, 2012; DCTerms, 2012) and the Data Documentation Initiative (DDI), which allows a best-fitting implementation with the applied software system Eprints (Eprints, 2013). The elements of Dublin Core and DDI have been used as a basis for the definition of a tailor-made metadata format, shown in Table 1.

TABLE 1: Metadata elements of the Clearing House of Transport and Mobility

Element	DCTerms	DDI		Element	DCTerms	DDI
General information on data set						
Title	title	titl		Subjects	tableOf Contents	topcClas
Title (abbreviation)		altTitl		Website for data set	relation	othrStdy Mat
Summary, purpose of data collection	description	abstract		Possible links to other data sets	references	
Keywords	subject	keyword		Resource ID	identifier	
Temporal and spatial information						
Time period of data collection	temporal	timePrd		City, municipality	spatial	geogCover
Country	spatial	geogCover		Additional information		notes
Federal state	spatial	geogCover				
Investigation and data collection methodology						
Data set variables		dataDscr		Data collection tool		collMode
Investigation design		timeMeth		Population		universe

Frequency of data collection	frequency	frequenc		Analysis unit		anlysUnit
Data collection method		dataColl		Sample size		sampProc
Access to data set						
Collected by	contributor	data Collector		Data format(s)		fileType
Client	creator	authEnty		Size of data set		dimensns
Contact for data set	publisher	producer		General comments to data set		notes

3. Metadata Concept of the MONITOR Database

The data repository within the MONITOR portal operated by the Institute of Air Transport and Airport Research (FW) was built in the years 2009-2011 as part of the EU-funded project MONITOR (Monitoring System of the Development of Global Aviation). The objective of this project was to build a permanent monitoring system and a corresponding data repository for the identification and evaluation of global long-term trends with regard to the aviation sector. Within the MONITOR project, the repository was elaborated in cooperation with an Advisory Committee and based on a stakeholder survey aimed at experts from the field of the aviation industry, politics, NGOs and society. Within this, the most important step was the design of a metadata scheme. Due to its high level of awareness and the advantages through accepted standardisation, an orientation on the Dublin Core Standard (DC, 1998) was chosen as starting point. Further metadata elements were added with special attention to the nature of the data and statistics to be described (Kokus et al., 2010; Kokus et al., 2011).

For example, the tailor-made MONITOR metadata elements that are presented in Table 2 include six specific elements dedicated to temporal aspects in order to describe data sources with regard to the time frame they cover, their temporal resolution and publication intervals (i.e. "temporal coverage", "time horizon", "temporal resolution", "frequency of publication", "date of release" and "temporal scope of publication"). It is thus possible, for instance, to classify a forecast study in detail with regard to the time frame the corresponding data covers, to distinguish further between pre-defined short-, medium- and long-term time ranges and to address the temporal resolution of the included data (e.g. yearly, monthly). In addition, the user is provided with information on the frequency of publication and the date of release in the case of a unique publication, or the temporal scope of publication if it is a publication series.

The following list of MONITOR elements partly but significantly differs from those of the Clearing House shown in section 2. For instance, while the Clearing House only hosts data sets, MONITOR also contains scientific publications on air transport. Here information about authors has to be collected. Therefore, a formatted type "person" was defined in order to cover all relevant information about the author (e.g. name, address, affiliation, etc.). This is a good example for showing the challenge of merging two different data formats into one common one. Solutions, not only regarding the common data format but also the presentation of search results and filters, had to be found for a lot more mismatches and differences.

TABLE 2: Metadata elements of the MONITOR data repository

Element	DC	New		Element	DC	New
A. Category Information						
Resource type	type			Content category		☑
B. Basic Information						
Title	title			Subject/keywords	subject	
Publisher/supplier	publisher			Online link	relation	
Author	creator			Description	description	
C. Technical Information						
Spatial coverage	coverage			Temporal scope of publication	date	

Temporal coverage	coverage			Data status		☑
Time horizon		☑		Data format	format	
Temporal resolution		☑		Size of data file		☑
Variables		☑		Language	language	
Frequency of publication		☑		Resource identifier	identifier	
Date of release	date					
D. Information for Using the Data (Usefulness, Possibilities and Limitations)						
Availability	available			Target group		☑
Method of data collection/processing		☑		Users of the data		☑
Collection reason		☑		Relation	relation	
Collection financer		☑		Alternative data sources		☑
Field(s) of application		☑		Possibilities and limitations of the data		☑

4. Common Metadata Concepts for the Interlinked Transport Statistics Portal STRADA@DLR

With the different metadata schemes of both DLR data portals presented here, the organisational challenge for the creation of STRADA@DLR was to establish a consensus as to which metadata information is necessary for the STRADA portal and should be presented. In addition, it had to be clarified which metadata elements of both institutes have a similar or identical meaning and can be merged at all. For some elements, harmonisation and title adaptations were unavoidable in order to make a consideration possible for implementation with regard to STRADA.

These processes included a detailed discussion among MONITOR and the Clearing House about the definition and understanding of the metadata elements which were already used in the existing data repositories. For example, the metadata attribute "spatial coverage", which is the title of one of the STRADA metadata elements (see Table 3), had to merge the detailed geographical description used by the Clearing House (with three sub-elements for country, federal state and city/municipality) with the description of the MONITOR portal that uses one global attribute – called "spatial coverage" – and the agreement was to find the lowest common denominator. Similar agreements were also found with regard to the merging of other metadata elements (e.g. attribute "person" in section 3).

Finally, a decision was to be made as to which elements should appear at the search result teasers of the targeted STRADA@DLR portal. Due to common experiences, FW and VF agreed to provide users at this point with short and most important metadata information only. Therefore, only the elements "title", "subject/keywords" and "description" per each data entry will be presented to give users a quick orientation as to whether the data source is appropriate. Another important aspect when designing the portal was the selection of search filters or facets based on metadata elements. It was decided to choose three filters that allow differentiating data, and which search results according to categories and spatial as well as temporal coverage. Filters and results teasers are discussed in more detail in section 5.

5. Technical Implementation of the Common Metadata Scheme

The technical implementation of the common metadata scheme was driven by the underlying search framework used to realise the STRADA@DLR portal. For this portal the same framework as for the MONITOR portal was used. This tailor-made framework called "KnowledgeFinder" stores all metadata directly attached to each data set. For this purpose "KnowledgeFinder" uses the open source software "Apache Subversion" (SVN). SVN is a software versioning and revision control system (Subversion, 2013). With this software it is possible to annotate single files with arbitrary metadata using so-called SVN properties. These properties were used to store the content of the common metadata scheme. For this purpose "KnowledgeFinder" applies

JavaScript Object Notation (JSON) format to represent the common elements inside an SVN property (Crockford, 2006). The advantage of this SVN-based solution is the simple mapping of metadata to a corresponding data set. Another important point to mention is the powerful metadata versioning mechanism of the underlying revision control system. It is planned to use this mechanism in future portal versions. One possibility could be to generate a version-based provenance out of the STRADA metadata.

To realise a search over both metadata sources, conversion and mapping between the different schemes had to be established. Due to the fact that the MONITOR portal metadata is already available within the SVN repository, only this metadata had to be mapped to the common scheme. In the case of Clearing House metadata, in a first step all data sets had to be imported into an SVN repository. The challenge of this SVN import was to establish a common file naming scheme to ensure a consistent handling of the data set creation, update and deletion. For this purpose, unique IDs of the Clearing House repository are used to identify a single data set. If a data set consists of multiple files, a zip archive is used to subsume all files in a single one. In a next step all relevant Clearing House metadata are harvested and converted into the JSON specific format. This conversion is performed with the help of the XPath Query Standard (Clark et al., 1999). An XPath query allows a simple addressing of single nodes inside a XML tree. With such queries the relevant content of each metadata attribute is extracted from the XML-based metadata. After the conversion a mapping of each metadata attribute with respect to the common metadata scheme was performed. Table 3 shows this mapping in detail. For instance, the Clearing House element “spatial coverage” is derived from the three elements “country”, “Federal State” and “municipality”. Then there is temporal coverage, which is mapped from two different formats and stored as simple strings. The common elements “Frequency of publication” and “Frequency of data collection” are solely mapped from one data source as they don’t have any appropriate counterpart. As a last step the mapped metadata elements were attached as an SVN property to the corresponding SVN file that finally holds the data set.

TABLE 3: Metadata elements of STRADA@DLR including the mapping of both portals

Element	Clearing House	Monitor	Search result teaser	Search filter	Metadata information sheet
Title	Title	Title	☑		☑
Publisher	Client	Publisher/supplier			☑
Author	Collected by	Author			☑
Subject/keywords	Keywords	Subject/keywords	☑		☑
Description	Summary, purpose of data collection	Description	☑		☑
Spatial coverage	Country / Federal State City / municipality	Spatial coverage		☑	☑
Frequency of publication	-/-	Frequency of publication			☑
Temporal coverage	Time period of data collection	Temporal coverage		☑	☑
Frequency of data collection	Frequency of data collection	-/-			☑
Time period of data collection	Time period of data collection	Temporal scope of publication			☑
Link	[determined]	[determined]			☑
Data category	Subjects	Content category		☑	☑

The “KnowledgeFinder” search engine indexes all data sets and their common metadata. A user interface is provided which allows a keyword-based search over this metadata repository. As mentioned above (see section 4) some metadata elements are used as search filters to narrow down the search results. Because of the different use in the original data repositories, the attribute "spatial coverage" has a special two-step filtering mechanism. Figure 1 shows an example of this filter. First, the user has to select a high-level spatial coverage (a). As shown, the user can choose between a world-wide, continental-wide, country-wide and region-wide spatial coverage. When the user selects one item, the search engine filters out all data sets with the appropriate high-level spatial coverage (b). This two-step filtering allows easy access to the different kind of spatial coverage types and limits down the number of presented metadata filters. This filter is realised with the help of region lists. Such lists are the basis for determining the spatial coverage (world, continental, country or region-wide) and, thus, allow the generation of the shown two-step mechanism. To identify countries, for example, the ISO 3166 country codes list is used (ISO, 1997).

Figure 2 shows another example for the filters based on the metadata elements "temporal coverage" and "data category". With the help of the "temporal coverage" filter, the user can select data items within a certain temporal range. To realise this ranged filter, the temporal coverage strings are transformed during the indexing process into an integer range of years. Meanwhile, the "data category" filter shows four examples for the common eight content categories established between the two institutes. Again, to build these filters a mapping had to be performed between the original metadata categories and the common metadata categories.

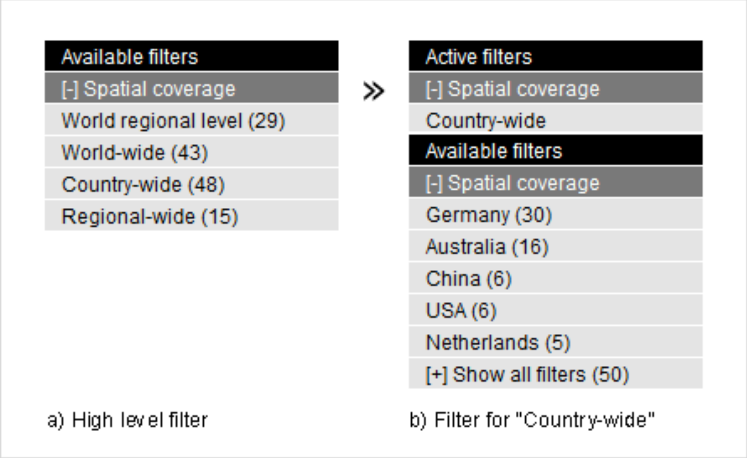

FIG. 1. Two-step filtering based on the metadata element "spatial coverage"

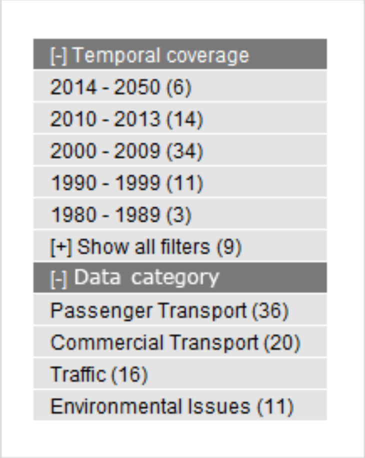

FIG. 2. Filter for the metadata elements "temporal coverage" and "data category"

6. Conclusion

As presented in this project report, a huge effort was put into the task to realise a common transport statistics and data search portal for creating additional value for transport researchers through an easily accessible data repository with modern search facilitates and detailed metadata descriptions. Given the special condition that both DLR institutes/portals involved had already independently deployed and approved data repositories with own tailor-made metadata schemes in operation, the challenge was to agree on a common metadata scheme besides the existing ones. In this situation, the orientation on Dublin Core, which was in any case already being used as the basis for the individual metadata schemes of both institutes, was an advantage. The thus-generated degree of standardisation facilitated a common understanding, helped to clarify the concrete meanings of the existing metadata elements and made it easier to agree upon the definition and technical specification of the new metadata elements for STRADA@DLR, which will be launched soon.

However, severe difficulties and obstacles had to be overcome as the aforementioned portals were developed independently and following different foci and approaches. While the Clearing House data portfolio is limited to transport and mobility statistic data sets, the MONITOR portal hosts a more diverse variety of data sets, publications, studies and prognoses. To find one common data format to describe elements of both portals was the biggest challenge and needed intensive discussions and, finally, many compromises in search reliability. Once the STRADA portal is operational, more investigation on search statistics is needed in order to fine-tune the format, its merging, applied search routines and filters. Besides that, it is also planned to perform a user acceptance study in order to evaluate the performance of the search engine.

A next step will be to integrate further DLR portals on transport data and information. As the portals most likely to be integrated are rather technically oriented, the integration of technical data set descriptions will bring another layer of complexity into the STRADA portal. Again, the STRADA data format has to be adjusted; format merges have to be discussed, new search filter and facets have to be defined and integrated into the STRADA system.

Participating partners of DLR research institutes did not know of any operational portal combining different transport data repositories while using Dublin Core Standards. A corresponding literature study in the field of transport research has been performed. Thus, no best practices or published expertise could be used and referenced.

References

Clark, James and DeRose, Steve. (1999). XML Path Language (XPath) Version 1.0. W3C Recommendation. Retrieved April 5, 2013, from http://www.w3.org/TR/xpath/.

Crockford, Douglas. (2006). RFC 4627: The application/json media type for JavaScript Object Notation. Retrieved April 5, 2013, from http://tools.ietf.org/html/rfc4627.

DC. (1998). Dublin Core Metadata Element Set, version 1.0. Retrieved April 5, 2013, from http://www.dublincore.org/documents/1998/09/dces/.

DC. (2012). Dublin Core Metadata Element Set, version 1.1. Retrieved April 5, 2013, from http://www.dublincore.org/documents/dces/.

DCTerms. (2012). DCMI Metadata Terms. Retrieved April 5, 2013, from http://dublincore.org/documents/dcmi-terms/.

Eprints. (2013). Official Website. Retrieved April 5, 2013, from http://www.eprints.org/.

ISO. (1997). ISO 3166-1 Country Codes. Retrieved June 21, 2013 from http://www.iso.org/iso/country_codes

Kokus, Petra, Ralf Berghof, Michael Hepting, Alexandra Leipold, Alfons Schmitt, Ab Hoolhorst and Zoltan Bilacz. (2010). D3: Report on network of sources and data description. Unpublished project report. Cologne: German Aerospace Center.

Kokus, Petra, Ralf Berghof, Michael Hepting, Alexandra Leipold and Alfons Schmitt. (2011). D 9: The MONITOR data base. Unpublished project report. Cologne: German Aerospace Center.

Subversion. (2013). Apache™ Subversion: Enterprise-class centralised version control for the masses. Retrieved April 5, 2013, from http://subversion.apache.org/.