Project Acronym: thinkMOTION
Grant Agreement number: 250485
Project Title: Digital Mechanism and Gear Library goes Europeana

D6.2 - Intermediate report on entered metadata sets
Revision: 1.1

Authors:
Claudia Schneider (RWTH Aachen)
## Revision History

<table>
<thead>
<tr>
<th>Revision</th>
<th>Date</th>
<th>Author</th>
<th>Organisation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>01.06.2011</td>
<td>C. Schneider</td>
<td>RWTH</td>
<td>Draft</td>
</tr>
<tr>
<td>1.0</td>
<td>18.06.2011</td>
<td>C. Schneider</td>
<td>RWTH</td>
<td>Final version</td>
</tr>
<tr>
<td>1.1</td>
<td>20.06.2012</td>
<td>S. Falke</td>
<td>IUT</td>
<td>Review</td>
</tr>
</tbody>
</table>

**Statement of originality:**
This deliverable contains original unpublished work except where clearly indicated otherwise.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.
Contents

1. Introduction and Workpackage objectives................................................. 4
2. Resources, workflow and results.................................................................. 4
   2.1 Resources for metadata ........................................................................ 5
   2.2 Workflow ......................................................................................... 5
   2.3 Entered metadata sets into thinkMOTION database ........................... 5
ANNEX 1: Simplified Metadata Editor for Patent Documents.......................... 7
1. Introduction and Workpackage objectives

This document provides information on the results of WP6 “Entering metadata for content” which have been achieved during the 2nd year of thinkMOTION project. The aim of the work package is to provide an adequate set of metadata for each content item. In general, metadata are defined as follows:

“Metadata describes other data. It provides information about a certain item's content. For example, an image may include metadata that describes how large the picture is, the colour depth, the image resolution, when the image was created, and other data. A text document's metadata may contain information about how long the document is, who the author is, when the document was written, and a short summary of the document.”

Main tasks

The work in WP6 “Entering metadata for content” includes two tasks:

The objective of Task 6.1 is to enter an appropriate set of metadata for each content item. The task involves the completion and (if necessary) the correction of the metadata entered as a result of the workflow in WP3. The metadata sets for the different types of content are defined in accordance with the standards of Dublin Core (DC) and Europeana Semantic Elements (ESE). Additionally, the language of the entered metadata sets is marked in order to facilitate the translation work in WP8.

Task 6.1 also interacts with WP7 “Who-is-who”, where the rudimentary metadata set for persons which was defined in the beginning (see “Guidelines for entering metadata” in D6.1), has been refined for the final set of metadata for biographies/persons (see D7.1).

The aim of Task 6.2 is the preparation and publication of guidelines for entering metadata. These guidelines cover the conversion of existing metadata and the entering of new metadata in line with the required metadata for Europeana. In addition, the guidelines are the basis for the staff training activities in WP10.

2. Resources, workflow and results

The following chapter describes shortly the resources for metadata and the applied workflow for entering metadata within the thinkMOTION project. A detailed description is given in Deliverable D6.1. The focus will be on the results achieved by the consortium during the 2nd year of the project.

---

http://www.techterms.com/definition/metadata
2.1 Resources for metadata

In D6.1 resources for metadata were described, which the partners had identified during the 1st project year, such as university libraries, archives, bibliographic catalogues and patent databases for text based contents. These sources also delivered metadata for the 2nd project year.

2.2 Workflow

Within the thinkMOTION project all partners are working on the web-based database “ProDB”. This database supports an efficient workflow for the registration and maintenance of all metadata assigned to each content item by offering appropriate input masks (see D6.1).

As described in Deliverable D6.1, several tools and tutorials were developed during the first year of the project.

Another tutorial has been developed recently, which describes the entering of metadata for patent documents in detail. This was necessary, because the metadata for patent documents differ from those of other text documents and so the categories of the input masks had to be adapted. The result is shown in Figure 1; the tutorial is attached in ANNEX 1.

![Figure 1: Simplified Metadata Editor for Patent Documents](image)

Especially the tutorials have contributed to a fast adaption of all partners to the necessary steps and workflows. Therefore, the work during the 2nd year of thinkMOTION could proceed without or with only minor problems.

2.3 Entered metadata sets into thinkMOTION database

Compared to the results achieved in the first year of the project, the consortium has reached a significant progress in the amounts of metadata sets, which is shown in Table 1.

Especially the number of registered metadata for images has been increased successfully; the same applies for text documents. Almost 70% of these text documents are articles (from journals, proceedings, etc.) and other documents. The category “other” covers mainly teaching materials, such as lecture notes and examinations tasks.

The consortium has entered about 18,000 metadata sets for images, which were obtained from different sources like books, patents or articles but also from external content providers such as museums.
Beside this analogue content, the amount of multimedia items has been increased to almost 1.200 metadata sets for animations and movies. Within the 2\textsuperscript{nd} year of thinkMOTION project 72 new mechanism descriptions have been created. In D6.1 a detailed description of ProDB was given, especially regarding the special tool “mechanism description”. This function allows the creation of very valuable items mainly for expert users of DMG-Lib. The creation of these items is of course desirable, but at the same time very time-consuming and requires a lot of knowledge about mechanism science.

Table 1: Entered metadata sets during 1\textsuperscript{st} and 2\textsuperscript{nd} project year

<table>
<thead>
<tr>
<th>Type of content</th>
<th>Amount of entered metadata sets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1\textsuperscript{st} year</td>
</tr>
<tr>
<td>Text documents</td>
<td>5.987</td>
</tr>
<tr>
<td>Article</td>
<td>2.276</td>
</tr>
<tr>
<td>Book</td>
<td>274</td>
</tr>
<tr>
<td>PhD</td>
<td>83</td>
</tr>
<tr>
<td>Patent</td>
<td>1.689</td>
</tr>
<tr>
<td>Other</td>
<td>300</td>
</tr>
<tr>
<td>Persons</td>
<td>4.972</td>
</tr>
<tr>
<td>Images</td>
<td>140</td>
</tr>
<tr>
<td>Movies</td>
<td>0</td>
</tr>
<tr>
<td>Animations</td>
<td>0</td>
</tr>
<tr>
<td>Mechanism descriptions</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total (without persons)</strong></td>
<td><strong>6.127</strong></td>
</tr>
</tbody>
</table>

The numbers shown in Table 1 represent the status “metadata entered” and do not give information on the amount of items which are available online for DMG-Lib and Europeana users. This is due to the applied workflow within the thinkMOTION project. The first step of item production is entering metadata, only then the process of scanning and processing starts. After that, the content is uploaded and finally prepared for online presentation. This process leads to the mentioned delay between the work steps. The actual amount and development of online available items is reported in the documents related to WP5. Also the numbers given for “persons” do not give a statement on the expected result of 1.500 biographies. The exact figures are given in the reports on WP7.
ANNEX 1: Simplified Metadata Editor for Patent Documents
TUTORIAL

Simplified Metadata Editor for Patent Documents

Version 1.0
Ilmenau, 2012-02-08

Author: Veit Henkel (IUT)
# Table of Contents

1. Preamble ........................................................................................................................................ 1  
2. Short Explanation of Terms used in *ProDB* Masks and Tutorial ........................................... 1  
3. Fields of Input Mask for Patent Documents ............................................................................. 2  
4. Workflow for Entering a Patent Document into *ProDB* ............................................................ 4  
5. Some Examples for Entering Patents into *ProDB* ..................................................................... 7
1. Preamble

For entering metadata of documents into the DMG-Lib database ProDB, which is used in the thinkMOTION project, a set of input masks, called Simplified Metadata Editor for Common Documents, was developed.

Patent documents have some metadata which are different from common documents and so it was necessary to develop some special input masks for patent documents. This set of input masks is called Simplified Metadata Editor for Patent Documents and is based on the Simplified Metadata Editor for Common Documents.

In this tutorial only the Simplified Metadata Editor for Patent Documents is described. But for a better understanding, it is necessary to read the tutorial for the Simplified Metadata Editor for Common Documents first. Especially adding of new persons (inventors of the patent) is not described in this tutorial, because it is similar to the workflow for common documents. But this step is absolutely necessary if a patent should be entered into DMG-Lib database ProDB and the inventor is not yet stored in the database.

Please note: The important keywords of the step descriptions in this tutorial are in bold letters. The names of buttons or fields, which are visible in the English ProDB version, are in italics.

2. Short Explanation of Terms used in ProDB Masks and Tutorial

For newer patents, standards for the notation of the bibliographical data in the patent documents are used in most of the important industrial countries. Unfortunately these standards have changed over the years and the older patents have often their own numbering systems depending on time and country. That makes it not easy for the thinkMOTION project staff.

The main idea of collecting patents in the DMG-Lib is not to build a database, which can be used for an extensive and complete patent research similar to the databases of the patent offices! Our aim is it to collect interesting patents in the field of motion systems, without the claim to completeness over all patents and without all the research functionality of the established patent databases. That allows us in the thinkMOTION project to reduce the full bibliographical set of metadata of the patents to a minimum set of metadata for the DMG-Lib.

Important metadata are the Title of the invention (patent document), the Inventor and the Date of Publication of the current patent document. The process of the patent issuance has more than one step and that is the reason for more than one date (day, month, year) at the patent documents. Normally, the date of publication of the current document should be used for the DMG-Lib. Further metadata are the Country Code, which tell us the country in which the patent was published and the so called Kind-of-Document Code (Kind-Code).

During the process of patent issuance, more than one document for the same patent will be produced. Among other things they differ usually from each other by the Kind-Code. A simplified example for German patents after the year 2004: The letter “A” in the Kind-Code means: this unexamined document was applied and was laid open for inspection in this current form and “B” means: this document is the patent specification which was granted in this current form.
The used characters for the Kind-Code can reach from “A” to “Z” and are normally followed by a number usually for the version. The content of these documents, which belongs to the same invention (patent), can be changed not or only a little and the title and the inventors are normally the same. So we should import only one of these documents. If available, please entering the patent specification document in the latest version. If not available, e.g. the patent was not granted, take the patent application document. In the DMG-Lib, the Kind-Code is only important to have a unique identifier for the patent documents.

The next metadata is the International Patent Classification number (IPC). The IPC is a hierarchical system of terms, which indicates the subject to which the invention relates and gives the user further details of the document’s content. In some cases, especially for older patents or for patents of some countries the IPC is not available. The information source for the IPC numbers can be the metadata set in the patent offices’ database or the patent document itself. Older patents are often post-classified, so that the IPC is only entered in the patent database and is not written in the patent document.

Why are the IPC numbers important for DMG-Lib respectively thinkMOTION? In the patents, technical terms are often paraphrased with words which nobody is using and nobody knows. The IPC numbers gives the patent documents a kind of keywords in a common and understandable technical language. The patent offices provide IPC-lists in different languages. These lists contain an assignment between the IPC number and the classification keywords. (E.g. IPC = ‘B66F 3/04’: “Devices, e.g. jacks, adapted for uninterrupted lifting of loads with racks actuated by pinions with several racks”)

In the thinkMOTION project, we are going to use the IPC numbers to import the IPC-terms (“keywords”) from these lists automatically. This saves you the time for searching and entering the IPC terms manually. The advantage of these IPC-terms or keywords in the DMG-Lib is that the user gets the possibility to find the document with his or her common technical language even though these terms are not written in the full text of the patent document.

For newer patent documents, normally all bibliographical data are marked in the patent document with the so-called INID-Codes. INID stands for Internationally agreed Numbers for the Identification of (bibliographic) Data. These numbers are helpful to find the correct metadata in the patent documents and to avoid faults. The INID-Codes are printed in the patent document nearby the metadata normally in brackets or in a circle. Unfortunately, the INID-Codes are not used in any patents, especially older patents or patents from particular countries. So that it is not a help in any case. In the Simplified Metadata Editor for Patent Documents the fields of the input masks are marked with the INID-Codes to give a support for entering metadata.

It is not necessary to know all details about patent codes. More information, give the standards of the WIPO - World Intellectual Property Organization (www.wipo.int). But it is not necessary for the thinkMOTION project.

3. Fields of Input Mask for Patent Documents
In this chapter the fields for entering metadata for patent documents into ProDB are shortly described. The Workflow, or how these fields are to be filled in, is described in the next chapter.
Figure 1 shows the input mask for adding a new metadata set for a patent document, which is not yet in the ProDB. Table 1 gives a short description of the input fields, the related INID-Codes and the necessity for entering into ProDB.

Table 1 - Meaning of the metadata input fields in the ProDB (see Figure 1)

<table>
<thead>
<tr>
<th>Metadata input field in ProDB</th>
<th>Description</th>
<th>INID-Code</th>
<th>Necessity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Country Code</td>
<td>Mostly 2 characters in front of the publication number which characterize the country in which the patent document was published (Examples: EP, US, ES, IT, FR, DE, RO, …) Hint: You can pin the Country Code if you entering many patents from the same country</td>
<td>19</td>
<td>Must be filled if known</td>
</tr>
<tr>
<td>Publication Number and Kind Code</td>
<td>The number of the patent document without the country code and followed by the “Kind-of-document code” according WIPO-Standard ST.16; it indicates the type of the document (Examples: A1, B1, …)</td>
<td>11, 13</td>
<td>Must be filled if known</td>
</tr>
<tr>
<td>Title</td>
<td>Title of the invention as written in the document. If translations are available, please add a further title fields by pressing the plus button</td>
<td>54</td>
<td>Mandatory (must be filled)</td>
</tr>
<tr>
<td>Language code of the title (no field name visible in mask)</td>
<td>The code must be entered according to ISO 639-1 language codes, which can be found in the online help for ProDB handling.</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Inventors</td>
<td>Name(s) of inventor(s)</td>
<td>72</td>
<td>Must be filled if known</td>
</tr>
<tr>
<td>Date of publication</td>
<td>Date of making this document available to the public – not other mentioned dates such as the date of filing the application (INID code 22). Even though the invention was made before, the date of publication of the current document has to be typed into this field!</td>
<td>4x^2</td>
<td>Mandatory (must be filled)</td>
</tr>
<tr>
<td>IPC Number(s)</td>
<td>International Patent Classification number(s), which helps to characterize the content of the patent document; type in only the numbers not the keywords (we intend to add the keywords later automatically)</td>
<td>51</td>
<td>Must be filled if known</td>
</tr>
<tr>
<td>Comment</td>
<td>If you have internal comments for yourself or other</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
DMG-Lib colleagues, you can enter your notes in the field Comment. Please write all comments in English.

External metadata

External metadata are the original metadata which can be found for example in reference lists of papers, on websites of library databases, etc. Please always use copy & paste to fill in this field.

Hint: You can enter the external metadata here first and then use it for copy and paste to the other fields above.

1) INID - Internationally agreed Numbers for the Identification of (bibliographic) Data according to the WIPO-Standard ST.9 [http://www.wipo.int/standards/en/pdf/03-09-01.pdf]; WIPO - World Intellectual Property Organization; for newer patent documents, the INID-Codes are written at the document

2) INID-category 40, depends on the document type and the history of the patent application process, “x” stands for one of the number 1-8 (e.g. 45 means: making a patent document available to the public by printing or similar process on which grant has taken place on or before the said date)


For entering patent documents into ProDB, please use the Simplified Metadata Editor for Patent Documents, which is linked from our ProDB startup page (see Figure 2).

To avoid entering the same patent document several times, you have to check whether the patent is already in the ProDB or not. This is the same step as for common documents.

For this step you have two search fields, the Search for the publication number and the Search for a title of the patent document (see Figure 3).

The publication number search uses different information, the Country Code, the Document Number and the Kind-Code. (These terms are described above). This type of search is only successful if the fields for these metadata are filled for the already entered patent document.

But in a lot of cases, the publication number is only entered in a ProDB’s field for a parallel title of the patent document. (That is a mistake, which should be fixed.) In order to avoid double documents, it is more safety to use the search for the patent title instead of the search for the publication number!
While typing in the patent title, the ProDB automatically suggests similar titles which are already stored in the database. By moving the mouse cursor over the titles, additional metadata are overlaid in a yellow tool tip to support the decision if the current document already exists in the ProDB or not. For more details about this auto completion function, please read the tutorial of the “Simplified Metadata Editor for Common Documents”.

If the patent is not already in the ProDB, a new data set must be generated and an input mask (see Figure 4) appears by pressing the Add new dataset button.

The title from the search field will be taken automatically into the Title field of the input mask to save time. But please make sure that the title is correct. Then type in the abbreviation for the language of the title. If there are written titles in different languages in the patent document or in the patent office’s database, please add parallel titles by pressing the + button and type in these titles also. (Adding parallel titles is detailed described in the tutorial for the Simplified Metadata Editor for Common Documents.)
The next field is for the **Inventors**. Please type in the name of each inventor in a separate field and check separately whether the person is already stored in the ProDB. For more details, see the tutorial of the *Simplified Metadata Editor for Common Documents*. Please observe that in a patent document usually more persons are mentioned. **Please type in only the names of the inventors not of the applicant(s), the grantee(s), holder(s), assignee(s) or owner(s), attorney(s) or agent(s) or any other persons!**

The next field is the **Date of publication**. Normally, there is more than one date written in a patent. Please use the date of the publication of the current document, which you hold in your hand or you see at your screen. Do not use publication dates of other documents, which are related to the current document. Please work carefully to find the correct publication date of the current patent document.

The **IPC Number(s)** are the next fields in the input mask of the *Simplified Metadata Editor for Patent Documents*. Please find the IPC number at the patent document or in the patent offices' database.

Internal comments for yourself or for other thinkMOTION staff can be entered in the field **Comment**. Write your comments in English please to make it readable for others.

**External metadata** means the original text of the metadata how you have found it in literature, lists of authors, etc. Please fill in this field by Copy&Paste. External metadata usually include additional information and can be used to solve errors. Never write any own comments in this field.

After checking all your data input, please press the **Send new data** button. After this, a new metadata set for a patent document was generated. Changes can be made on this data set by using the search and *editSimple* function in the ProDB.
5. Some Examples for Entering Patents into ProDB

**Footnotes:**

(*) Type in the language codes of the titles

(**) Don’t use this date! It isn’t the publication date of the patent.
Footnotes:

*) In some cases the country code is not explicit written in the patent document and the document code is to take from the WIPO-Standard ST.3 In the example, the headline „EUROPEAN PATENT SPECIFICATION“ says that this document is an European patent. And that means, that the country code is „EP“.

**) Select the language of the title

***) Don’t use this date! It isn’t the publication date of the patent.
**Footnotes:**

*) In some cases the country code is not explicit written in the patent document and the document code is to take from the WIPO-Standard ST.3. In the example, the headline “United States Patent Office” says that this document is an European patent. And that means, that the country code is “US”.

**) Select the language of the title

****) Don’t use this date! It isn’t the publication date of the patent.

*****) Don’t use this number! It is the application serial number of the patent not the patent publication number!
In some cases the country code is not explicit written in the patent document and the document code is to take from the WIPO-Standard ST.3. In the example, the headline „Rumania says that this document is an Rumanian patent. And that means, that the country code is „RO”. **) Select the language of the title ***) Don’t use this date! It isn’t the publication date of the patent.