

Establishing and Delivering XML Exports – Guidelines for EFG Content Providers

Version: 1.1
21/06/2010

Authors:
Franco Zoppi (Consiglio Nazionale delle Ricerche – Istituto di Scienza e Tecnologie
dell'Informazione), Francesca Schulze (Deutsches Filminstitut - DIF)

Contents

| | |
|---|---|
| 1 Preliminary Remarks | 3 |
| 2 Establishing and Delivering XML Exports | 4 |
| 3 EFG Ingestion Service | 5 |
| 4 EFG System Functional Outline | 6 |
| 5 Terms & Definitions | 7 |

1 Preliminary Remarks

These guidelines are dedicated to inform you, as an EFG content provider, how to establish and deliver XML exports to the EFG system. They inform you about several possible delivery methods and the EFG ingestion process from a technical point of view. If you are interested in which of your local metadata elements (a.k.a. database fields) are relevant for the EFG database and hence should be included into your local XML export(s) for EFG please read the document: "Preparing Metadata for XML Exports – Guidelines for EFG Content Providers".

Once established, we kindly ask you to send your XML export(s) together with all relevant information concerning your content delivery to the EFG data co-ordination team at the German Film Institute - DIF:

- | | | |
|----------------------|--|--------------------|
| a) Francesca Schulze | f.schulze@deutsches-filminstitut.de | +49 69 961 220 701 |
| b) Julia Welter | welter@deutsches-filminstitut.de | +49 69 961 220 403 |

We will validate your XML export(s) against the EFG content delivery criteria and write a proposal on how your local database schema can be mapped to the EFG metadata schema. This means that we will allocate the elements from your local export to those of the EFG schema and write the transformation rules. Once you approved your mapping we will forward it together with your XML exports to our IT-partner Consiglio Nazionale delle Ricerche – Istituto di Scienza e Tecnologie dell'Informazione (CNR-ISTI). The team at CNR-ISTI will set up the import filter and ingest your contribution into the EFG database. Before your data go online you have the possibility to validate the representation of your data in the EFG web portal in a content checker tool. The same procedure applies for the publication of your contribution in the Europeana portal. We would like to mention that we will not publish your contribution before we received your final approval for this.

If you have any further questions regarding your content contribution or the ingestion process please do not hesitate to contact us.

The following documents should be read in conjunction with these guidelines:

- [EFG metadata schema \(textual version\)](#)
- [EFG metadata schema \(tabular overview\)](#)
- [EFG metadata schema \(XSD file\)](#)
- [Preparing Metadata for XML Exports – Guidelines for EFG Content Providers](#)

2 Establishing and Delivering XML Exports

Please consider the following guidelines when you provide content and (meta)data for the EFG System:

1. The export must be provided using an UTF-8 encoding. This ensures that local special characters (å, ø, etc.) are transferred correctly in the EFG database and eventually will be displayed correctly in the EFG web portal. Please do not use the ISO-8859-1 or other encoding schemes. For further reading please visit the following link:
<http://en.wikipedia.org/wiki/UTF-8>.
2. The local exports should not contain syntactical errors. The XMLStartlet tool can be used to check the syntax of the XML files: <http://www.ibm.com/developerworks/library/x-starlet.html>.
3. Each record / data set in the export files must contain a unique identifier. Such an ID is usually automatically generated by your local database system(s), and must remain unchanged.
4. The XML files must not be edited manually. They must be established through an automated procedure - e.g. export function supplied by a DBMS¹.
5. Please provide different XML files for each kind of data you are delivering, i.e. for film works, persons, digital objects (in EFG called Item) etc.. The XML exports listed hereunder contain a number of sample records from the German Film Institute (DIF):
 - Digital object record: [sample_dif_item000.xml](#)
 - Film work records: [sample_dif_fw0000.xml](#)
 - Person records: [sample_dif_p000.xml](#)
6. A file size limit must be considered: It is fixed at 500 KBytes or to the next higher multiple of each single entry dimension (e.g. 128 KBytes * 5 = 512 KB). As a consequence, if that size is exceeded, multiple files should be generated with a simple naming convention like – e.g. – DIF_Persons0000.xml, DIF_Persons0001.xml, etc..
7. If a provider needs to change the XML export structure on a later stage, a new XML containing the new local schema must be supplied in advance. This XML file will be checked by the Ingestion Service for its consistency and correctness.

The following guidelines are optional:

8. In addition to the XML files containing the actual (meta)data, also an XSD file containing the local schema should be provided if possible. The local schema is used by the ingestion service to check the consistency of the ingested (meta)data. For an example please see the film works schema export from the German Film Institute (DIF): [sample_dif_schema.xsd](#)
9. The preferred method to transfer metadata to EFG is to set up a web access to your XML files. Based on such an access, the EFG Ingestion Service can check the existence of new metadata and content on a regular basis. We suggest to supply an HTTP access (and subject

¹ Database Management System

to that OAI-PMH, RSS, or any protocol/interface even proprietary that partners would use), or alternatively FTP. HTTPS should be excluded to avoid authentication management issues.

If you intent to set up an OAI-PMH interface to make your (meta)data harvestable on a regular basis please read the guidelines on the EFG project website:

<http://www.europeanfilmgateway.eu/gs1.php>.

3 EFG Ingestion Service

The EFG Ingestion Service is designed to harvest (meta)data from the content providers' repositories and to ingest it into the EFG Storage according to the EFG schema specification given in the [EFG metadata schema \(XSD file\)](#).

The Ingestion Service supports the:

- Ingestion of content into EFG from local repositories via an XML export
- Implementation of the EFG Schema
- Translation of metadata from the local schemata to the EFG Schema
- Storage of EFG metadata in the Limbo of the Storage Service via OAI-PMH

This figure bellow illustrates the several functional parts of the EFG Ingestion Service:

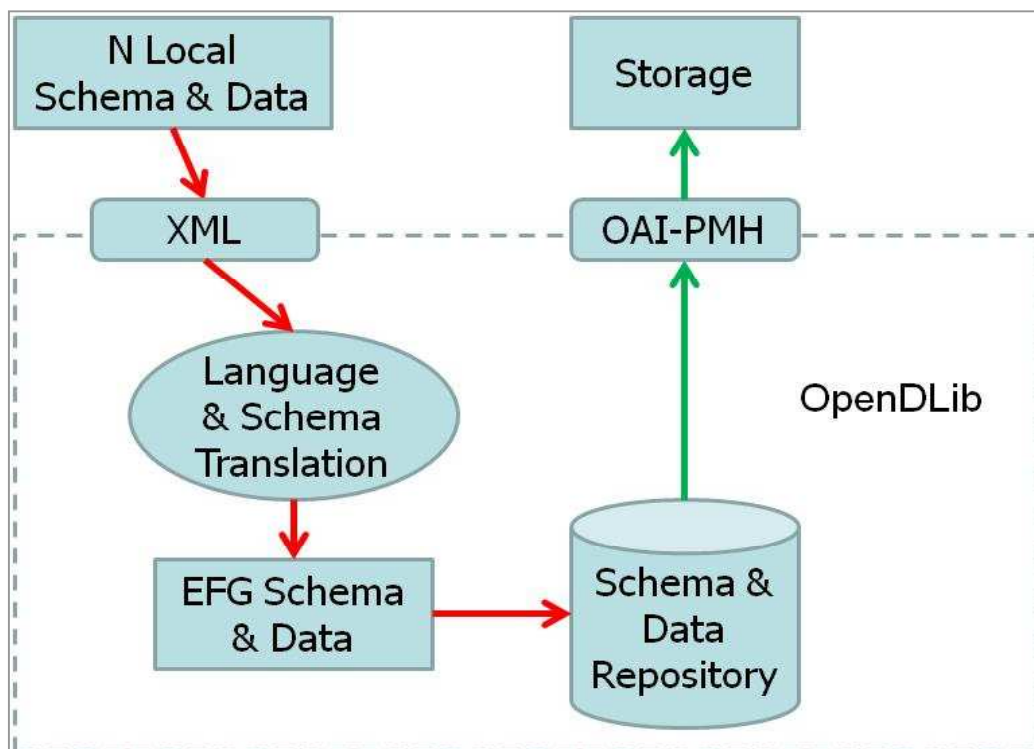


Figure 1 – Ingestion Service Functional Schema

4 EFG System Functional Outline

This figure shows how the Ingestion Service is embedded in the entire EFG System structure.

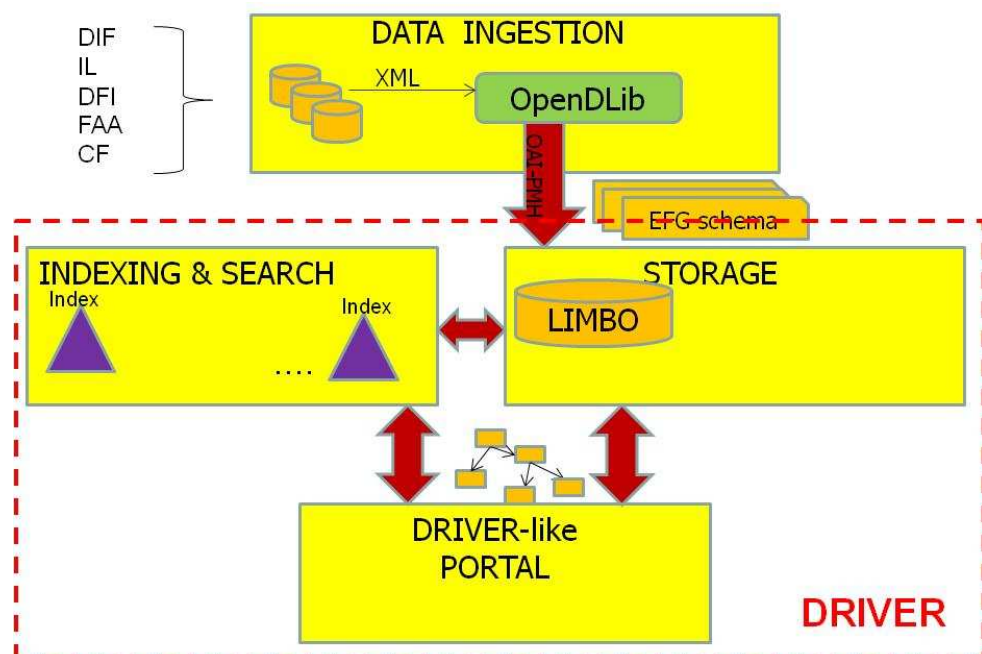
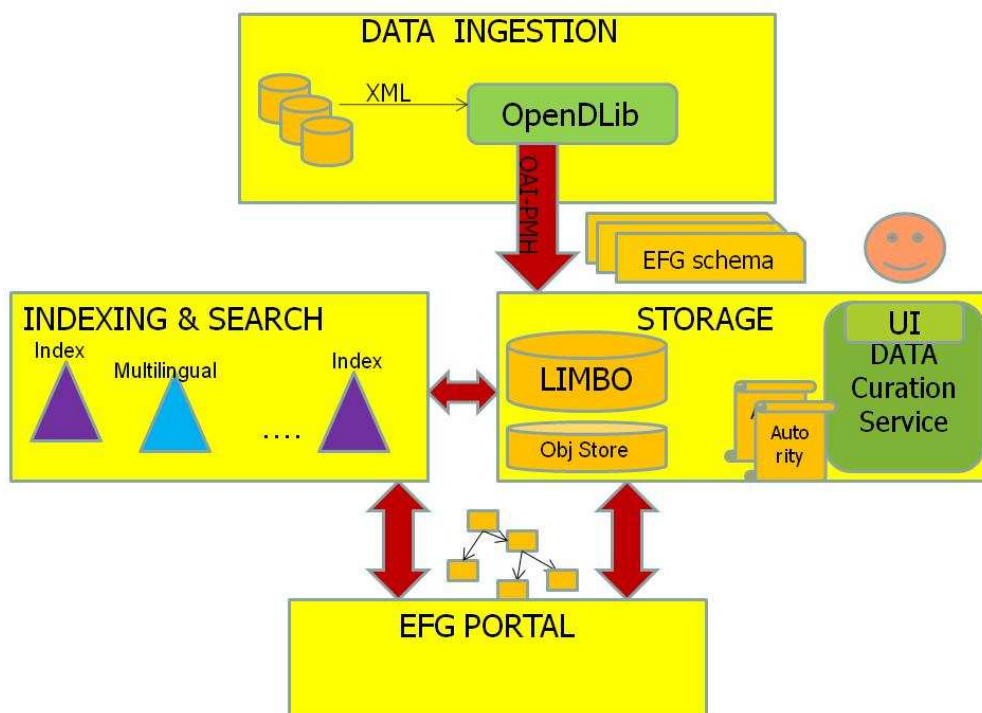


Figure 2: Functional Outline of the EFG System

If you are further interested in the EFG system please read the document [D 2.3 Adaptation of DRIVER software components](#) provided on the EFG project website (for project partners only) or contact the EFG Data Co-ordination team at the German Film Institute (DIF).

5 Terms & Definitions

| Term | Definition |
|---------------------|--|
| EFG metadata schema | The schema for (meta)data representation defined by EFG |
| Limbo | The Storage sub-component where (meta)data are stored before being “checked and approved” by data curators |
| Local schemata | The schemata for (meta)data representation used by the content providers |
| OAI-PMH | Open Archive Initiative - Protocol for Metadata Harvesting |
| Storage | The EFG system component which implements the Limbo, Obj Store, Data Curation, Authority File Control & Metadata Editor Services |
| Content | Content is digital information with a reference to an individual object of the real world or is born digital. Examples: photographs, videos, letters, censorship documents, etc. |
| Metadata | Metadata are “data about data” which are extracted from the content providers’ local databases. They describe content and can be produced by authority files or controlled vocabularies. Examples: filmographic data, temporary and spatial data, etc. |
| Content provider | Institution that delivers content and metadata to the EFG system. |
| DBMS | Database Management System |