**Open Data Management for Public Automated Translation Services**

From ITS

Jump to: [navigation](https://www.w3.org/International/its/wiki/Open_Data_Management_for_Public_Automated_Translation_Services#mw-navigation), [search](https://www.w3.org/International/its/wiki/Open_Data_Management_for_Public_Automated_Translation_Services#p-search)

Editors:

* Dave Lewis, CNGL at Trinity College Dublin – dave.lewis@cs.tcd.ie
* Felix Sasaki, DFKI / W3C Fellow - felix.sasaki@dfki.de, fsasaki@w3.org
* Asun Gomez-Perez, Universidad Politécnica de Madrid - asun@fi.upm.es
* Serge Gladkoff, GALA CRISP – sgladkoff@gala-global.org

Draft Version 0.4; 16th June 2014

**Contents**

 [[hide](https://www.w3.org/International/its/wiki/Open_Data_Management_for_Public_Automated_Translation_Services)]

* [1 Introduction](https://www.w3.org/International/its/wiki/Open_Data_Management_for_Public_Automated_Translation_Services#Introduction)
	+ [1.1 Purpose and context of this document](https://www.w3.org/International/its/wiki/Open_Data_Management_for_Public_Automated_Translation_Services#Purpose_and_context_of_this_document)
	+ [1.2 Who should read this document](https://www.w3.org/International/its/wiki/Open_Data_Management_for_Public_Automated_Translation_Services#Who_should_read_this_document)
	+ [1.3 How and when to contribute](https://www.w3.org/International/its/wiki/Open_Data_Management_for_Public_Automated_Translation_Services#How_and_when_to_contribute)
* [2 Terminology](https://www.w3.org/International/its/wiki/Open_Data_Management_for_Public_Automated_Translation_Services#Terminology)
* [3 Social/Business Goals](https://www.w3.org/International/its/wiki/Open_Data_Management_for_Public_Automated_Translation_Services#Social.2FBusiness_Goals)
* [4 Interoperability Goals](https://www.w3.org/International/its/wiki/Open_Data_Management_for_Public_Automated_Translation_Services#Interoperability_Goals)
* [5 Data Management Lifecycle Reference Model for Automated Translation](https://www.w3.org/International/its/wiki/Open_Data_Management_for_Public_Automated_Translation_Services#Data_Management_Lifecycle_Reference_Model_for_Automated_Translation)
* [6 Interoperability Requirement Recommendations](https://www.w3.org/International/its/wiki/Open_Data_Management_for_Public_Automated_Translation_Services#Interoperability_Requirement_Recommendations)
	+ [6.1 General Data Management Requirements](https://www.w3.org/International/its/wiki/Open_Data_Management_for_Public_Automated_Translation_Services#General_Data_Management_Requirements)
	+ [6.2 Bitext Data Management Requirements](https://www.w3.org/International/its/wiki/Open_Data_Management_for_Public_Automated_Translation_Services#Bitext_Data_Management_Requirements)
	+ [6.3 Lexical-Conceptual Data Management Requirements](https://www.w3.org/International/its/wiki/Open_Data_Management_for_Public_Automated_Translation_Services#Lexical-Conceptual_Data_Management_Requirements)

**Introduction**

**Purpose and context of this document**

This document assemble recommendations related to public automated translation services. The recommendations are based on interoperability experience and knowhow from various projects and initiatives: the [LT-Web project](http://www.w3.org/International/multilingualweb/lt/); the [MONNET project](http://lemon-model.net/); the [FALCON project](http://www.falcon-proejct.eu/); the [LIDER project](http://www.lider-project.eu/); the [QTLaunchpad project](http://www.qt21.eu/launchpad/) and the [interoperability research at CNGL](http://www.cngl.ie/research/research-areas/interoperability-and-analytics/).

**tbd: purpose help to formulate reqdocs. then: mention projects.**

The document is being developed through public consultation, including via: the [ITS Interest Group](http://www.w3.org/International/its/ig/); the [Open Linguistics WG](http://linguistics.okfn.org/) at the Open knowledge Foundation; the [Linked Data for Language Technology](http://www.w3.org/community/ld4lt/) Community Group; the [Best Practice in Multilingual Linked Open Data](http://www.w3.org/community/bpmlod/) Community Group and the [OntoLex](http://www.w3.org/community/ontolex/) Community Group.

The aim of the document is to inform the development of public automated translation services about requirements for achieving interoperability in data management, which is a key aspect of realizing such services. The document should provide answers to questions like:

1. What type of interoperability requirements are relevant for public automated translation services?
2. What goals can be achieved by following these requirements?
3. How does a reference model for data management in automated translation services look like? To this end, the document adopts a *data management lifecycle* approach to structuring the requirements of different stakeholders.

**Who should read this document**

**tbd: cef out of focus.**

In Europe, the [Connecting Europe Facility](https://ec.europa.eu/digital-agenda/en/connecting-europe-facility) is a European funding program to be exectued between 2014-2020. It encompasses the development of public automated translation services. One audience of this document is to inform CEF planning about interoperability requirements for automated translation.

The document does not focus on CEF Automated Translation services alone, though with a [call for preparatory actions pending (section 3.1.7)](https://ec.europa.eu/digital-agenda/sites/digital-agenda/files/WP2014%20-%20official%20published.pdf) this is a timely issue. More broadly this document also aims at formulating data interoperabilty requirements for automated translation services in general: the document should be relevant for all such services. Two examples of relevant communities and efforts are given below:

1. The [Wikimedia Foundation](http://wikimediafoundation.org/) has stated at the [MultilingualWeb workshop 2014](http://www.multilingualweb.eu/documents/2014-madrid-workshop/2014-madrid-program) (see [slide 10](http://www.w3.org/International/multilingualweb/2014-madrid/slides/sharma.pdf)) that machine translation infrastructure is relevant technology for fostering multilingual content generation in Wikipedia. The requirements formulated in this document can provide input for developing machine translation infrastructure in Wikipedia.
2. More and more machine translation services are available on the Web, being closely integrated with core Web technologies like HTML5. Although such services are mostly developed by non public organisations, i.e. by private companies, following interoperability requirements can improve the quality and broad acceptance of the services.

**How and when to contribute**

The editors invite your feedback by email and will advance this work via the ITS Interest Group at the W3C. The aim is to have a stable version of the document ready by the mid July 2016.

**Terminology**

The document uses the following terminology. Some terms refer to each other; the order below should enable the reader to understand these relations easily.

1. **Public automated translation services**. Public automated translation services are services that provide automated translation functionality, for example to be used by European public administration to deliver content to the public of multilingual Europe.
2. **Sharing of language resources**. Automated translation services use language resources (LR), e.g.: parallel corpora of translated text for machine translation training purposes, translation memories, machine readable lexica etc. By sharing of LR we mean the re-usage and distribution of such resources, with explicit legal information about licensing of resources.
3. **Standards**. Standards are commonly accepted data structures and practices designed for public use, to be used when defining the requirements with interoperability build in for technical building blocks of data and technology to handle language resources and for creating machine translation services. This document advocates for open standards in language industry. Example standards are:
	* [TMX](http://www.gala-global.org/oscarStandards/tmx/) for representing translation memories;
	* [XLIFF](https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xliff) for interchange of localisable information and related metadata.
	* ITS 2.0 for metadata annotation of content
	* ISO, ASTM, W3C specifications and public recommendations
	* RDF for linked data
4. Several aspects are important related to standards:
	* This document does not recommend a closed set of standards, but rather the best practice of using standards for implementing automated translation services (see below).
	* Standards are relevant on a technical level in various areas (see below), but also as a means to fulfil legal requirements (e.g. the forehand mentioned sharing of language resources).
	* Key technical areas that benefit from standards are:
		+ Representation of parallel information to be used in automated or manual translation services, so-called **bitext**. If a translation direction is given one refers to **source language** and **target language**.
		+ Representation of content to be translated, e.g. Web content
		+ Definition of software interfaces for creating and accessing automated translation services
		+ Representation of additional information to improve the quality of automated translation services, e.g. related to lexcial / conceptual information, terminology, provenance etc. Such additional information is referred to as **metadata**. The anchoring of metadata in the actual information is referred to as **annotation**.
		+ Guidelines for post-editing, that is manual correction of machine translation output.
5. **Open standards**. Open standards are standards that are developed and maintained through an open, international consultative process and that can be implemented on a royalty-free basis, that is, without any licensing requirements.
6. **Interoperability**. Standards foster interoperabilty between developers and users of automated translation services, that is: they lower the cost and increase quality and speed of the services.
7. **Curation of the language resource**. Language resources sometimes are not available in a standardised form. To ease their application in an interoperable manner, they need to be converted into standardised formats or not to be made accessible via standardised APIs.
8. **Data management**. Handing of data is a key aspect of automated translation services. Data here means language resources as well as the content to be translated itself, data about users, about provenance of language resources and translated content etc.
9. **Language Quality Assurance**. Repeatable, feasible and sustainable production process to obtain the statistically valid measure of human perception of the quality of the translated material and its measure of satisfaction to the target project requirements. The practical LQA methods to assess language quality of public material are to be taken into account, e.g. MQM global error typology, and also methods of assessment such as described in ASTM work item ASTM WK46397, “Development of a complete methodology, including a simplified quality metric, for crowd-sourced expert language quality assessment targeted at nonprofit web sites and other documents of public interest”.

**Social/Business Goals**

These recommendation aim to satisfy the following social and business goals:

* Encourage the rapid and sustainable sharing of language resources to enable predicatable progress in improving the availability, cost and quality of public automated translation services.
* To respect the ownership of language resource data and support the transparent processing of this data in accordance with the licensing conditions specified by those owners.
* To reduce the cost of achieving widespread data interoperability between parties involved in the deployment and use of public automated translation services and to reduce the cost related data management activities within those parties.

**Interoperability Goals**

These recommendations also aim to satisfy the following technical interoperability goals:

* Provide best practice for data management as input to procurement specifications, for example for the CEF automated translation service and future European Language Cloud, as well as proprietary language clouds that are now emerging.
* To promote the adoption of best practice in open, Web based content and data management.
* To enable the active curation of the language resource data required by public automated translation services. This involves the continuous and systematic collection and quality assurance of parallel text, target language models and multilingual lexical-conceptual data based on the human quality judgments associated with the deployment and use of public automated translation.
* To ensure presence of interoperability data and service features enabling full and meaningful engagement of language service and language technology communities, as well as necessary prerequisites for the professional and public feedback and participation.

**Data Management Lifecycle Reference Model for Automated Translation**

The recommendation in this document are based on the following model data management for automated translation. The model assumes three main groups of data to be taken into account:

1. The content being translated with assistance of automated translation.
2. The bitext used in training automated translation components.
3. The lexical-conceptual resources associated with the source and target that assist in the consistent use and translation of terms.

The Reference Model identifies some high-level activities which are judged important to the data management related to automated translation, and which should therefore be subject to data interoperability recommendations. This set of activities are not intended to be fully comprehensive, as many variations of these process chains will potentially use automated translation, however they are sufficient to highlighting the main interoperability issues to be addressed.



**@@please, change “Trans QA” to “LQA”, or “Linguistic QA”.**

**Interoperability Requirement Recommendations**

These requirement recommendations focus on the data management lifecycles identified in the reference model. They should be complemented by additional inoperability recommendation on the services required to implement these activities.

**tbd: structure of requirements.**

Requirement are marked as mandatory (M) or optional (O).

**General Data Management Requirements**

1. (M) Public automated translation services should use open standards for representing content, metadata and annotations. “Open” means that the standards are available on a royality-free basis, to maximize re-use and adoption.
2. (M) All data used and content processed by public automated translation services should be annotated with license information in a machine readable format.
3. Where licensing terms permit, content and data should be published on the web and dereferencable via a unique URI.
4. (O) To enable content and data to be used as a resource in the generation of automated translation engines they should be:
	* Annotated with common machine readable meta-data to allow them to be automatically indexed and discovered. This meta-data should conform to a profile of the DCAT meta-data.
	* Provided with a persistent URL.
5. (M) It should be possible for third parties (general public, individual experts and language service providers alike, as well as automated language services) to submit error, QA or corrective annotations to published data, provided it is presented in a common format, with metadata conformant to one of the commonly accepted and documented universal error typologies, and/or appropriate quality metrics.
6. (O) The status of submitted error, QA or corrective annotation, as it is considered for integration by the curator of the data set, should be made available with reference to the original submissions and in a common format.

**Bitext Data Management Requirements**

1. (O) If content that is translated with assistance of public automated translation services is published on the web together with the source content, then the corresponding bitext aligned at a segment level should also be published. A different license to the source and target language documents may be used.
2. (O) Bitext data may be discovered via a web-based API that can return specific segment bitext data selected via query parameters. Parameters may include:
	* Source and target languages.
	* Presence of terms or phrases.
	* Translation provenance meta-data including: identification of the automated translation component used and its operational parameters; characteristics of post-editors; characteristics of the postedits (edit type, edit distance, time to postedit); characteristics of the QA (quality assurance) method applied (parameters an d assessment guidelines); QA annotations; and annotation links to specific lexical-conceptual resources.

**Lexical-Conceptual Data Management Requirements**

1. (O) If content is translated with assistance of the public automated translation in combination with lexical-conceptual data and is published on the web together with the source content, then the source and/or target content should also be available in a form that annotates the relevant terms or multi-word units with the lexical-conceptual concepts used.
2. (O) Lexical-conceptual data may be discovered via a web-based API that can return lexical and conceptual data selected via query parameters. Parameters may include:
	* Terms or multi-word units in source or target languages.
	* Contextual information for the terms or multi-word units
	* Conceptual restrictions upon which to filter results
	* Lexical restrictions upon which to filter results
	* Provenance meta-data of the lexical-conceptual data sought, including: source of the data; process by which the data was created; current status of data still under curation.