KCONNECT

www.kconnect.eu

Initial Data Management Plan

Deliverable number	*D6.1*
Dissemination level	*Public*
Delivery date	*23 July 2015*
Status	*Final*
Author(s)	*Allan Hanbury, Célia Boyer, Jon Brassey, Richard Dobson, Ljiljana Dolamic, Pavel Pecina, Konstantin Pentchev, Angus Roberts, Agata Rukat*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644753 (KConnect)

Executive Summary

This deliverable is the initial version of the data management plan. In this document, the data generated by the KConnect project is identified and the current status of the data management, archiving, preservation and licensing plans are given. In particular, this initial analysis of the data indicates where further efforts are required to clearly specify these aspects of the data management plan. The final version of the data management plan is D6.3 due in July 2016.

Each section of this deliverable describes a data resource identified in the KConnect project. The format followed in each section corresponds to the structure proposed in the European Commission Guidelines on Data Management in Horizon 2020 [1].

The following table summarises the data resources presented:

Name	**Description**	**License**
Knowledge Base	The knowledge base is a warehouse of semantically integrated data sets published originally by third parties. It includes information on drugs, drug targets, drug interactions, diseases, symptoms, adverse events, anatomies and imaging modalities. In addition to the data sets it includes link sets that map data between the different data sets and/or provide semantic relationships.	Depends on individual datasets
Khresmoi Summary Translation Test Data 1.1	This dataset contains data for the development and testing of machine translation of sentences from summaries of medical articles between Czech, English, French, and German. The original sentences are sampled from summaries of English medical documents crawled from the web in 2012 and identified to be relevant to 50 medical topics.	CC-BY-NC, version 3.0 unported
Khresmoi Query Translation Test Data 1.0	This data sets contains data for development and testing of machine translation of medical queries between Czech, English, French, and German. The queries come from general public and medical experts.	CC-BY-NC, version 3.0 unported
Text and annotation indices	The dataset comprises texts annotated and indexed by the KConnect semantic annotation pipeline, in order to create a searchable index with links to the KConnect knowledge base. The individual annotations and indices will be individually described in the next version of the deliverable.	Depends on the individual texts
HONSearchLogs	This data set contains the query logs collected from various search engines maintained by HON. The search engine logs are collected over a period of over 3 years (since November 2011) and are continuing to be collected.	CC-BY-SA, version 3.0 unported
Trip Database search logs	These logs record search terms and articles viewed. In addition this data is linked to the user so that information about profession, geography, professional interests etc. can be considered.	Currently no formal licensing information

The South London and Maudsley NHS Foundation Trust (SLAM) Hospital Records	The SLAM electronic health record (EHR) contains records for 250,000 patients in a mixture of structured and over 18 million free text fields. The Clinical Record Interactive Search (CRIS) application allows research use of the pseudonymised mental health electronic records data (with ethics approval since 2008).	Data access is governed through a patient led oversight committee
Region Jönköping County Patient Records	Test data consisting of 50 fictitious patient records with the same data structure and content as a real patient record, in Region Jönköping County's electronic health records system (Cosmic), would have.	Currently no formal licensing information

Table of Contents

1 Introduction 5
2 Knowledge Base 5
3 Summary Translation Test Data 6
4 Query Translation Test Data 7
5 Annotated Text Data 8
6 HON Search Logs 9
7 TRIP Database Search Logs 10
8 KCL Patient Records 11
9 Qulturum Patient Records 11
10 Conclusion 12
11 References 12

1 Introduction

This deliverable is the initial version of the data management plan. In this document, the data generated by the KConnect project is identified and the current status of the data management, archiving, preservation and licensing plans are given. In particular, this initial analysis of the data indicates where further efforts are required to clearly specify these aspects of the data management plan. The final version of the data management plan is D6.3 due in July 2016.

Each section of this deliverable describes a data resource identified in the KConnect project. The format followed in each section corresponds to the structure proposed in the European Commission Guidelines on Data Management in Horizon 2020 [1].

In summary, Sections 2 to 5 deal with data for which no privacy issues exist (knowledge base, machine translation training data, and annotations and indices), while Sections 6 to 9 deal with data in which care needs to be taken to ensure that privacy is preserved (search logs and medical records).

2 Knowledge Base

2.1 Name

Knowledge Base

2.2 Description

The knowledge base is a warehouse of semantically integrated data sets published originally by third parties. It includes information on drugs, drug targets, drug interactions, diseases, symptoms, adverse events, anatomies and imaging modalities. In addition to the data sets it includes link sets that map data between the different data sets and/or provide semantic relationships. The data is available as RDF and is loaded into a GraphDB [2] repository.

Original data sets:

- Drugbank
- UMLS
- RadLex
- DBPedia (medical subset)
- GeoNames

2.3 Standards and metadata

The data is available in different RDF formats: RDF-XML, NTriple, Turtle, TriG, TriX and RDF-JSON. It can be queried via SPARQL and the KB exposes the OpenRDF REST API.

2.4 Data sharing conditions

Data sharing varies according to the sharing conditions associated with the original data sets.

2.5 Archiving and preservation

Archiving and preservation varies according to the Archiving and preservation arrangements associated with the original data sets. Ontotext stores backups of the data sets converted to RDF and the corresponding link sets on its servers.

2.6 Licensing information

Licensing varies according to the licensing of the original data sets.

3 Summary Translation Test Data

3.1 Name

Khresmoi Summary Translation Test Data 1.1

3.2 Description

This dataset contains data for development and testing of machine translation of sentences from summaries of medical articles between Czech, English, French, and German. The original sentences are sampled from summaries of English medical documents crawled from the web in 2012 and identified to be relevant to 50 medical topics.

The original sentences in English were randomly selected from automatically generated summaries of documents from the CLEF 2013 eHealth Task 3 collection [1] which were found to be relevant to 50 test topics provided for the same task. Out-of-domain and ungrammatical sentences were manually removed. The sentences are provided with information on document ID and topic ID. The topic descriptions are provided as well. The sentences were translated by medical experts into Czech, French, and German and reviewed. The data sets can be used, for example, for the development and testing of machine translation in the medical domain.

3.3 Standards and metadata

The data is provided in two formats: plain text and SGML. They are split according to the section (dev/test) and language (CS – Czech, DE - German, FR - French, EN – English). All the files use the UTF-8 encoding. The plain text files contain one sentence per line and translations are identified by line numbers. The SGML format suits the NIST MT scoring tool. Topic description format is based on XML, each topic description (<query>) contains the following tags:

Tag	Description
<id>	topic ID
<discharge_summary>	reference to discharge summary
<title>	text of the query
<desc>	longer description of what the query means
<narr>	expected content of the relevant documents
<profile>	profile of the user

3.4 Data sharing conditions

Access to this data set is widely open under the license specified below.

3.5 Archiving and preservation

The data set is distributed by the LINDAT/Clarin project of the Ministry of Education, Youth and Sports of the Czech Republic and is available here: http://hdl.handle.net/11858/00-097C-0000-0023-866E-1

3.6 Licensing information

The data set is made available under the terms of the Creative Commons Attribution-Noncommercial (CC-BY-NC) license, version 3.0 unported. A full description and explanation of the licensing terms is available here: http://creativecommons.org/licenses/by-nc/3.0/

4 Query Translation Test Data

4.1 Name

Khresmoi Query Translation Test Data 1.0

4.2 Description

This data sets contains data for development and testing of machine translation of medical queries between Czech, English, French, and German. The queries come from general public and medical experts.

The original queries in English were randomly selected from real user query logs provided by Health on the Net foundation (750 queries by general public) and from the Trip database query log (758 queries by medical professionals) and translated to Czech, German, and French by medical experts. The test sets can be used, for example, for the development and testing of machine translation of search queries in the medical domain.

4.3 Standards and metadata

The data is split into 8 files, according to the section (dev/test) and language (CS - Czech, DE - German, FR - French, EN – English). The files are in plain text using the UTF-8 encoding. Each line contains a single query. Translations are identified by line numbers.

4.4 Data sharing conditions

Access to this data set is widely open under the license specified below.

4.5 Archiving and preservation

The data set is distributed by the LINDAT/Clarin project of the Ministry of Education, Youth and Sports of the Czech Republic and is available here: http://hdl.handle.net/11858/00-097C-0000-0022-D9BF-5

4.6 Licensing information

The data set is made available under the terms of the Creative Commons Attribution-Noncommercial (CC-BY-NC) license, version 3.0 unported. A full description and explanation of the licensing terms is available here: http://creativecommons.org/licenses/by-nc/3.0/

5 Annotated Text Data

5.1 Name

Text and annotation indices

5.2 Description

The dataset comprises texts annotated and indexed by the KConnect semantic annotation pipeline, in order to create a searchable index with links to the KConnect knowledge base.

There are several datasets, each held by the KConnect partner that is responsible for the underlying texts. In the next version of this deliverable, these datasets will be individually identified and described, as the plan for the progress of the KConnect work does not allow this to be done at this stage.

5.3 Standards and metadata

Texts are annotated using a Text Encoding Initiative (TEI) compliant framework, GATE [3, 4], to create documents encoded with UTF-8, in GATE XML format.

Annotations are linked to the knowledge base using URIs, and are searchable using SPARQL

5.4 Data sharing conditions

Data sharing varies according to the sharing conditions associated with the underlying text collection.

5.5 Archiving and preservation

Archiving and preservation varies according to the archiving and preservation arrangements associated with the underlying text collection.

5.6 Licensing information

Licensing varies according to the licensing of the underlying text collection.

6 HON Search Logs

6.1 Name

HONSearchLogs

6.2 Description

Search Engine Logs provided by the Health On the Net Foundation (HON). This data set contains the query logs collected from various search engines maintained by HON. The search engine logs are collected over a period of over 3 years (since November 2011) and are continuing to be collected.

The search engine logs contain following information:

- query term
- users' IP address – which enables determining the geographical distribution of the search
- exact date and time of the query
- language
- information on the search engine used to perform the search (honSearch, honSelect, …)
- information on the link followed

6.3 Standards and metadata

The search logs will be provided in the XML format, for which the metadata will be provided. An illustration of the format draft is given in the Figure 1.

```
<LOG engine=honSearch>
    <IP>52.0.*.*</IP>
    <usertrack>52.0.*.*.*</usertrack>
    <time>24/Jun/2015:00:28:14 +0200</time>
    <query language='en' action='search' docType='html'>
        diabete
    </query>
    ...
</LOG>
```

Figure 1. Search Log format draft

6.4 Data sharing conditions

This data set is provided by HON for the project partners. This data can be used for analysis of users' behaviour linked to the search engine usage.

With the goal of preservation of the users' personal data, the original content of the search logs is modified by HON. This modification consists of masking the part of the users' IP address, however keeping the parts of the IP which would enable the analysis of the global users' whereabouts.

In the above shown format draft the alternations of the original query logs are marked with "*".

6.5 Archiving and preservation

The original search logs are archived and kept on HON premises for the period of 5 years. These archives consist of the original, non-treated search logs. Investigation is underway for a possibility for longer-term preservation of the anonymised logs.

6.6 Licensing information

The HONSearchLogs will be made available on demand by the partners. The data are distributed under the terms of the Creative Commons Attribution-ShareAlike (CC-BY-SA), version 3.0 unported. A full description and explanation of the licensing terms is available here: https://creativecommons.org/licenses/by-sa/3.0/

7 TRIP Database Search Logs

7.1 Name

Trip Database search logs

7.2 Description

As users interact with the Trip Database (https://www.tripdatabase.com) the site captures the user's activity. It records search terms and articles viewed. In addition this data is linked to the user so that information about profession, geography, professional interests etc. can be considered. This may be useful in helping understand the search process, important documents, linked concepts etc.

There is considerable data going back multiple years and is constantly being collected.

7.3 Standards and metadata

There are no official standards.

7.4 Data sharing conditions

The data can be shared with the KConnect consortia with prior permission. Outside of KConnect the sharing of the data will be by negotiation.

Currently the data needs to be requested and downloaded by the Trip Database but an API is being considered.

7.5 Archiving and preservation

The data is stored on the Trip servers and these are backed up and saved on a daily basis. The production of the search logs is independent of the KConnect project and is increasingly core to the development of the Trip Database. As such the costs are seen as core to Trip.

7.6 Licensing information

There is currently no formal licensing information.

8 KCL Patient Records

8.1 Name

The South London and Maudsley NHS Foundation Trust (SLAM) Hospital Records

8.2 Description

The South London and Maudsley NHS Foundation Trust (SLAM) is the largest provider of mental health services in Europe. The hospital electronic health record (EHR), implemented in 2007, contains records for 250,000 patients in a mixture of structured and over 18 million free text fields.

At the NIHR Biomedical Research Centre for Mental Health and Unit for Dementia at the Institute of Psychiatry, Psychology and Neuroscience (IOPPN), King's College London we have developed the Clinical Record Interactive Search application (CRIS, http://www.slam.nhs.uk/about/core-facilities/cris), which allows research use of the pseudonymised mental health electronic records data (with ethics approval since 2008).

8.3 Standards and metadata

Through this model we will be able to provide access to a regular snapshot of the complete set of pseudonymised records in XHTML format.

8.4 Data sharing conditions

Records can be accessed by collaborators either onsite or through a remote secure connection.

8.5 Archiving and preservation

The record system is maintained by hospital IT services.

8.6 Licensing information

Data access is governed through a patient led oversight committee.

9 Qulturum Patient Records

9.1 Name

Region Jönköping County Patient Records

9.2 Description

Region Jönköping County (RJC) provides test data consisting of 50 fictitious patient records with the same data structure and content as a real patient record, in Region Jönköping County's electronic health records system (Cosmic), would have.

RJC will also provide real patient records if needed. In order to do so, the anonymisation process must be secured. The number of patient records that will be provided depends on how much data is needed in order to develop and test the KConnect solution. The first step of using test data instead of real patient records is necessary in order to secure that the tool is correctly adapted and implemented before real EHR content is used.

9.3 Standards and metadata

The use case(s) decided on will determine how data will be manipulated, e.g. addition of metadata/annotations.

9.4 Data sharing conditions

Records data be created, stored/archived, accessed, used and kept secure by Jönköping staff on site.

9.5 Archiving and preservation

An archiving and preservation plan is under development.

9.6 Licensing information

There is currently no formal licensing information.

10 Conclusion

This deliverable presents the initial version of the Data Management Plan for the KConnect project. It identifies data that is being collected in the KConnect project and allows missing information to be identified. An updated version of the deliverable that will include the information currently not available will be released as D6.3 in July 2016.

11 References

[1] European Commission, Guidelines on Data Management in Horizon 2020, Ver. 1.0, 11 December 2013, http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf, accessed : 10.7.2015

[2] Barry Bishop, Atanas Kiryakov, Damyan Ognyanoff, Zdravko Tashev, and Ruslan Velkov, OWLIM: A family of scalable semantic repositories, Semantic Web 2(1), 2011, 33-42.

[3] H. Cunningham, V. Tablan, A. Roberts, K. Bontcheva (2013) Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. PLoS Comput Biol 9(2): e1002854. doi:10.1371/journal.pcbi.1002854

[4] H. Cunningham, et al. Text Processing with GATE (Version 6). University of Sheffield Department of Computer Science. 15 April 2011. ISBN 0956599311.