MOVING

Deliverable 6.2: Data Management Plan

(2nd updated version: March 2019)

Chrysa Collyda, Vasileios Mezaris, Sabrina Herbst, Franziska Günther, Paul Grunewald, Thomas Köhler, Angela Fessl, Ahmed Saleh, Till Blume, Falk Böschen, Ansgar Scherp, Iacopo Vagliano, Markel Vigo, Tobias Backes, Peter Mutschke, Andrzej Skulimowski

29/03/2019

Work Package 6: Project management

TrainiG towards a society of data-saVvy inforMation prOfessionals to enable open leadership INnovation

Horizon 2020 - INSO-4-2015

Research and Innovation Programme

Grant Agreement Number 693092

Dissemination level	*PU*
Contractual date of delivery	*30/09/2016 for the original deliverable*
Actual date of delivery	*30/09/2016 for the original deliverable,* *30/03/2018 for the updated version,* *29/03/2019 for the 2nd updated version*
Deliverable number	*D6.2*
Deliverable name	*Data management plan*
File	*MOVING_D6.2_v3.0_updated201.doc*
Nature	*Report*
Status & version	*2nd Updated version, March 2019*
Number of pages	*61*
WP contributing to the deliverable	*WP6*
Task responsible	*CERTH*
Other contributors	*All*
Author(s)	*Chrysa Collyda, Vasileios Mezaris, CERTH* *Sabrina Herbst, Franziska Günther, Paul Grunewald, Thomas Köhler TUD* *Angela Fessl, KC* *Ahmed Saleh, Till Blume, Falk Böschen, Ansgar Scherp, Iacopo Vagliano ZBW* *Markel Vigo, UMAN* *Tobias Backes, Peter Mutschke, GESIS* *Andrzej Skulimowski, PBF*
Quality Assessors	*Angela Fessl, KC*
EC Project Officer	*Hinano SPREAFICO*
Keywords	*Data Management Plan*

Table of contents

Executive Summary

This deliverable presents the 2nd updated version of the Data Management Plan of the MOVING project. In particular, it describes in detail the adopted management policy for the datasets that have been collected, processed or generated by the project. The utilised approach: (a) identifies which data and how they will be exploited or made publicly accessible so as to maximise their reuse potential, (b) specifies how these data will be curated and preserved, to support their reuse, and (c) identifies any data that should not be made publicly available and measures to be taken for their safe-keeping.

The European Commission (EC) has defined a number of guidelines / requirements for maximising scientific data's reuse potential, via making them easily discoverable, intelligible, usable beyond the original purpose for which they were collected and interoperable to specific quality standards. Using these guidelines as a basis, we apply the methodology that is outlined in Section 1. According to this approach, for each dataset we specify: (a) its name (based on a standardised referencing approach), (b) its description, (c) the utilised standards and metadata, (d) the applicable data sharing policy and (e) the intended actions for its archiving and preservation. Further explanation regarding the information that needs to be considered and reported for each one of the above points is given in Sections 1.1. to 1.5. In Section 1.6 we addressed issues pertaining to persistence of data, access to restricted data and metadata, the treatment of sensitive data, the server security strategy, and the informed consent of human participants. Subsequently, based on the methodology in Section 1.1 to 1.5, Section 2 lists and describes the datasets of the MOVING project in a per-work-package-basis (Sections 2.1 to 2.5). It should be noted that WP4 is focused on the software development of the platform, thus is not expected to generate any dataset, or use datasets other than those already specified by the other work-packages. The concluding Section 3 briefly summarises the information reported in the deliverable.

In the 2nd updated document we discuss 53 different datasets, where 30 of them being preexisting and have been used in MOVING, and 23 others being new datasets and have been created and used during the last year of the project. Most of the dataset are publicly available, or at least will be available to third parties under certain conditions (e.g. CC licenses, upon request, etc…); only a few of the considered datasets will not be publicly available due to copyright restrictions or the need to respect data protection laws.

The MOVING Data Management Plan is a working document that evolves during the lifetime of the project. For this reason, although no formal update of the deliverable was foreseen, we have already produced an updated version of the Data Management Plan in accordance with the project's needs; this 2nd updated deliverable is available via the project's website.

Abbreviations

Abbreviation	Explanation
ACM	Association for Computing Machinery
API	Application Program Interface
ASCII	American Standard Code for Information Interchange
BIB	BibTeX file format
BIN	Binary file format
BSBM	Berlin SPARQL BenchMark
BTC	Billion Triple Challenge
CC	Creative Commons license
CHIME	Center for Information Mining and Extraction
CSV	Comma-Separated Values
DBLP	DataBase systems and Logic Programming
DCC	Digital Content Creation
DCNNs	Deep Convolutional Neural Networks
DERI	Digital Enterprise Research Institute
DMP	Data Management Plan
DyLDO	Dynamic Linked Data Observatory
FOAF	Friend Of A Friend
GFDL	GNU Free Documentation License

Abbreviation	Explanation
GDPR	General Data Protection Regulation
IPR	Intellectual property rights
JSON	JavaScript Object Notation
LDC	Linguistic Data Consortium
LOD	Linked Open Data
LUBM	Lehigh University BenchMark
MOOC	Massive Open Online Course
MPD	Million Playlist Dataset
MPEG	Moving Picture Experts Group
NIST	National Institute of Standards and Technology
NLM	National Library of Medicine
NTCIR	NII Testbeds and Community for Information access Research
NUS	National University of Singapore
OAI	Open Archives Initiative
OCLC	Online Computer Library Center
PDF	Portable Document Format
RCV1	Reuters Corpus, Volume 1
RDF	Resource Description Framework
RTF	Rich Text Format

Abbreviation	Explanation
SPARQL	SPARQL Protocol and RDF Query Language
SQL	Structured Query Language
TRECVID MED	TREC Video Retrieval Evaluation Multimedia Event Detection
TRECVID SIN	TREC Video Retrieval Evaluation Semantic Indexing
TSM	Tivoli Storage System
TSV	Tab-Separated Values
TXT	Text
WAV	Waveform Audio File Format
WK	Wolters Kluwer
W3C	World Wide Web Consortium
XML	eXtensible Markup Language

1 Applied Methodology

The applied methodology for drafting the Data Management Plan of the project was based on the guidelines of the EC[1] and the DMP online tool[2], which can be used for implementing such a plan in a structured manner via a series of questions that need to be clarified for each dataset of the project. According to these guidelines, the Data Management Plan of MOVING addresses the points below on a dataset-by-dataset basis, reflecting the current status within the consortium about the data that will be produced:

- Dataset reference and name.
- Dataset description.
- Standards and metadata.
- Data sharing.
- Archiving and preservation (including storage and backup).

A more detailed description of the information that is considered and reported for each one of these subjects is provided in the following subsections.

1.1 Dataset reference and name

For convenient referencing of the data that will be collected and/or generated in the project we had to define a naming pattern. A referencing approach that contains information about the WP that owns/uses the dataset, the serial number of the dataset and the title of the dataset is the following: *MOVING_Data_"WPNo."_"DatasetNo."_"DatasetTitle"*. According to this pattern, an example dataset reference name could be *MOVING_Data_WP1_1_UserGeneratedContent*.

1.2 Dataset description

The description of the dataset that will be collected and/or generated includes information regarding the origin (in case of data collection), nature and scale of the data, as well as details related to the potential users of the data. Moreover, the description clarifies whether these datasets are expected to support a scientific publication, while information on the existence (or not) of similar data and the possibilities for integration and reuse is provided.

[1] https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[2] https://dmponline.dcc.ac.uk

1.3 Standards and metadata

This section outlines how the data will be collected and/or generated, and which community data standards (if any) will be used. Moreover, it provides information on how the data will be organised during the project, mentioning for example naming conventions, version control and folder structures. For a detailed overview of the used standards the following questions were considered:

- How will the data be created?
- What standards or methodologies will be used?
- Which structuring and naming approach will be applied for folders and files?
- How different versions of a dataset will be easily identifiable?

In addition this section reports the types of metadata that is created to describe the data and aid their discovery. How this information is created/captured and where it is stored is also reported. The aspects below were examined for determining the necessary ways and types of generating and using metadata:

- How these metadata are going to be captured/created?
- Can any of this information be created automatically?
- What metadata standards will be used and why?

1.4 Data sharing

This section describes how the collected and/or generated data will be shared. For this, it reports on access procedures and embargo periods (if any), and lists technical mechanisms and software/tools for dissemination and exploitation/re-use of these data. Moreover it determines whether access will be widely open or restricted to specific groups (e.g. due to participant confidentiality, consent agreements or Intellectual Property Rights (IPR), while it outlines any expected difficulties in data sharing, along with causes and possible measures to overcome these difficulties. In case a dataset cannot be shared, the reasons for this are mentioned (e.g. ethical rules of personal data and privacy-related considerations, intellectual property and commercial interests). Last but not least, identification of the repository where data will be stored, indicating in particular the type of repository (institutional, standard repository for the discipline, etc.) is also performed. The questions below were studied for concluding to the most appropriate sharing policy for each dataset of the project:

- How these data are going to be available to others?
- With whom will be the data shared, and under what conditions?
- Are any restrictions on data sharing required (e.g. limits on who can use the data, when and for what purpose)?

- What restrictions are needed and why?
- What actions will be taken to overcome or minimise restrictions?
- Where (i.e. in which repository) will the data be deposited?

1.5 Archiving and preservation

The established data archiving and preservation policy defines the procedures that will be put in place for long-term preservation of the data. In particular it indicates how long the data will be preserved and what is their approximate end volume. It also outlines the plans for preparing and documenting data for sharing and archiving. In case of not using an established repository, the Data Management Plan describes the resources and systems that will be put in place to enable the data to be curated and used effectively beyond the lifetime of the project.

A set of questions that were considered for defining the archiving and preservation policy for the datasets of the project is given below:

- What is the long-term preservation plan for the dataset (e.g. deposit in a data repository)?
- Are there sufficient resources, including storage and other equipment, to carry out this plan, or are any additional resources needed?

1.6 Summary of changes in the present document

Compared to the first updated Data Management Plan of MOVING, delivered in March 2018, the present document introducing the following changes:

- Nine new datasets are introduced in WP2:
    - MOVING_Data_WP2_2_Movielens_1M
    - MOVING_Data_WP2_3_hetrec2011_lastfm
    - MOVING_Data_WP2_4_LibraryThing
    - MOVING_Data_WP2_5_Aminer_DBLP
    - MOVING_Data_WP2_6_Aminer_ACM
    - MOVING_Data_WP2_7_MPD
    - MOVING_Data_WP2_8_MOOC_Evaluation_1
    - MOVING_Data_WP2_9_MOOC_Evaluation_2
    - MOVING_Data_WP2_10_MOOC_ForumEntries
- Fifteen new datasets are introduced in WP3:
    - MOVING_Data_WP3_21_PubMed-2-IR
    - MOVING_Data_WP3_22_IREON-2-IR

- MOVING_Data_WP3_23_Wiki-page-views
- MOVING_Data_WP3_24_Twitter-Presidential-Election-2016
- MOVING_Data_WP3_25_ACL
- MOVING_Data_WP3_26_DeGruyter
- MOVING_Data_WP3_27_TimBL
- MOVING_Data_WP3_28_BSBM
- MOVING_Data_WP3_29_LUBM
- MOVING_Data_WP3_30_Wikidata[3]
- MOVING_Data_WP3_31_CORE_DB
- MOVING_Data_WP3_32_ZBWECONNISMetedataDataset
- MOVING_Data_WP3_33_LawAndRegulationsDataset
- MOVING_Data_WP3_34_LectureVideoFragmentationDataset
- MOVING_Data_WP3_35_NeLO

Furthermore, two datasets that were previously included in the data management plan deliverable (MOVING_Data_WP1_1_Innovations_in_scholarly_communications_study, and MOVING_Data_WP2_1_EYUserStudyResultKC) were removed in this latest edition because they were no longer relevant to MOVING.

In addition to these changes, and relating to all datasets described in the following Section 2, we would like to clarify some issues pertaining to persistence of data, access to restricted data and metadata, the treatment of sensitive data, server security, and informed consent of human participants (where applicable).

Persistence of storage is achieved by:

- Storing the datasets made publicly available by MOVING in Zenodo[4]. Zenodo is a strong supporter of open data in all its forms (meaning data that anyone is free to use, reuse, and redistribute) and takes an incentives approach to encourage depositing under an open license. Also accepts up to 50GB per dataset (accepts multiple datasets); and most important there is no size limit on communities accounts.
- Storing in servers of MOVING partner institutions (TUD, GESIS, ZBW, PBF) the datasets that are not made publicly available, or are made available to entities outside the MOVING consortium in a controlled way, due to privacy concerns or in order to comply with the

[3] The use of this dataset is planned also for WP2 in the context of semantic profiling and the recommender system.

[4] http://help.zenodo.org/

restrictions introduced when generating these datasets,. These servers are both protected and backed-up according to the strict institutional regulations of each of these partners; and, being institutional servers, they will continue to exist long after the end of the MOVING project.

- Maintaining the public MOVING website and the internal wiki on a server of the MOVING partner institution CERTH, which again is protected and backed-up according to the institutional regulations and will continue to exist long after the end of the MOVING project.

The coverage of metadata strongly depends on the data collected in the different datasets and cannot be unified. Typical metadata for data available on the MOVING platform are for example title, author, publication year, keywords etc.

Access to restricted data and metadata within the MOVING consortium is controlled by the owner of the corresponding dataset. Only partners that demonstrate the need to access one such dataset are provided access to it, and only for the purpose specified in their access request. This is in accordance with the data protection and ethics procedures of MOVING and in compliance with the applicable regulation, including the General Data Protection Regulation (GDPR).

Access to restricted data and metadata for entities outside the MOVING consortium, for those datasets where controlled sharing of them with external entities may be possible and is foreseen in the corresponding dataset tables of Section 2 of the present document, is performed in a case by case basis, according to the provisions made in the corresponding dataset tables (e.g. for the MOVING_Data_WP3_11_ZBWEconomicsDataset, MOVING_Data_WP2_4_LibraryThing, and MOVING_Data_WP3_27_TimBL by contacting and establishing an individual agreement with ZBW).

The allocation of resources, thus, the allocation of the datasets is solely the responsibility of the MOVING partners' institutions. This involves the not only the needs and priorities determining the most effective course of the dataset collection itself, but also to maximise the effective use of the dataset for the corresponding purpose. No dataset includes sensitive personal data; furthermore, concerning personal data in general, which may be included in datasets that are collected with the involvement of human participants (e.g. as a result of interviews), these data are immediately anonymised. Only anonymous data that cannot be used for identifying a participant are stored and used in MOVING. For the storage of digital datasets in servers of MOVING partner institutions (TUD, GESIS, ZBW, PBF, and CERTH), these servers adopt the server security measures dictated by the IT support department of each institution for their institutional computing infrastructure, and typically include firewalls, up-to-date anti-virus software and adopting all good practices that are expected for institutional IT infrastructure.

For all datasets that were created with the involvement of human participants, consent agreements (see Appendix) have been collected and are retained by the MOVING partners that conducted the corresponding experiments.

2 Datasets in MOVING

This section lists the datasets that have been created or collected for the needs of the MOVING project, grouping them per WP (with the exception of WP4, which is focused on the software development of the platform, thus it is not expected to generate any new dataset, or use datasets other than those already specified by the other work packages). Based on the methodology presented in Section 1, each dataset is defined by: (a) its name, (b) its description, (c) the used standards and accompanying metadata, (d) the applied data sharing policy, and (e) the adopted mechanisms for its archiving and preservation. As a key component for the creation and management of these datasets, data privacy issues will be closely monitored from the beginning of the project by the entire consortium and the project's Data Protection Officer (Mr. Robert Lorenz from TUD), to ensure that the collection, use and sharing of the data will not raise ethical concerns. Finally, we should stress that all our plans for the collection, retention, storage and sharing of datasets in the project, as described in the present document, do comply with the applicable national and EU legislations.

2.1 WP1 Datasets

Dataset name	**MOVING_Data_WP1_3 _Interviews**
Dataset description	The dataset consists of 9 anonymised interview transcripts with young researchers (PhD and Master students). The interviews consist of questions about: • The behavior towards research information searching and information management strategies. • The usage of online-based tools for research. • Training behavior and usage of online-based training services. The interviews are conducted because at the moment no similar data exists. The dataset is used for the requirement analysis of the platform (Task 1.1). This data may only be reused for research purposes under the data sharing conditions mentioned below.
Standards and metadata	The dataset consists of anonymised interview transcripts (doc or rtf files). The interviews are conducted as structured interviews.
Data sharing	Access to the interview dataset is restricted to the MOVING project partners to ensure privacy protection and no abuse of the data. Results of the empirical analysis are accessible to third parties via a poster (https://doi.org/10.5281/zenodo.

	1137640) and in deliverable D1.1.
Archiving and preservation	The interviews dataset is permanently stored on file servers of TUD. File servers are protected by applying the commonly used security measures for preventing unauthorised access and ensuring that security software is up-to-date with the latest released security patches. Preservation is ensured by regularly backups of the file systems. Backups are either conducted internally with the local administrator of the Media Center or with the help of a service partner within TUD, that offers free of charge backups with IBM Tivoli Storage System (TSM) including a guaranteed maximum 10-year time span to preserve this data. Additionally duration may be discussed with said partner.

Dataset name	**MOVING_Data_WP1_4_LabUserStudies**
Dataset description	Contains data collected from the user studies in the lab. The goal of such studies is to verify whether the requirements of the platform are met. In order to do so we collect: • Questionnaire responses • Observations • Performance metrics This dataset is key to accomplish Task 1.4 and verify whether users' expectations with the MOVING platform are met. Some other parties may find it useful for reproducibility and repeatability reasons. The data is analysed and discussed in D1.3.
Standards and metadata	This data is formatted as CSV, Markdown and MS-Office doc file.
Data sharing	Open to the consortium since its creation, and open to all once the corresponding papers have been submitted for publication; except for log data, which gets stored in TUD servers and cannot leave them due to the German Data Protection Law.
Archiving and	Data is uploaded into the project's Zenodo[5] profile. Also, it is stored on the data servers of the "Interaction Analysis and Modeling Lab" at http://iam-

[5] https://zenodo.org/collection/user-moving-h2020

preservation	data.cs.manchester.ac.uk/data_files. This server automatically backups its contents to another machine through a CRON job.

Dataset name	**MOVING_Data_WP1_5_FocusGroupInterviewTranscript**
Dataset description	The dataset consists of one focus group interview transcripts with six young researchers (PhD) conducted in May 2017. The transcript consists of information about: • The behaviour towards research information searching and information management strategies. • The usage of online-based tools for research. • Training behavior and usage of online-based training services. • Feedback to the first Mockups of the MOVING platform The dataset is part of the implementation of the use case on training young researchers (Task 1.3). This data may only be reused for research purposes under the data sharing conditions mentioned below.
Standards and metadata	The dataset consists of an anonymised focus group interview transcript (doc or rtf files).
Data sharing	Access to the interview dataset is restricted to the MOVING project partners to ensure privacy protection and no abuse of the data.
Archiving and preservation	The interviews dataset is permanently stored on file servers of TUD. File servers are protected by applying the commonly used security measures for preventing unauthorised access and ensuring that security software is up-to-date with the latest released security patches. Preservation is ensured by regularly backups of the file systems. Backups are either conducted internally with the local administrator of the Media Center or with the help of a service partner within TUD, that offers free of charge backups with IBM Tivoli Storage System (TSM) including a guaranteed maximum 10-year time span to preserve this data. Additionally, duration may be discussed with the service partner.

2.2 WP2 Datasets

Dataset name	**MOVING_Data_WP2_2_Movielens_1M**
Dataset description	Movielens is a collection of 1,000,209 anonymous ratings of approximately 3,900 movies made by 6,040 MovieLens users who joined MovieLens in 2000.
Standards and metadata	Users, ratings and movies are stored in DAT format.
Data sharing	The dataset is publicly available for non-commercial purposes[6].
Archiving and preservation	The dataset and analysis results are stored persistently at ZBW premises on protected servers. The servers are protected by established security measures for preventing unauthorised access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups. The archiving and preservation of this dataset are performed by the GroupLens Research Project at the University of Minnesota; MOVING has no involvement in this process.

Dataset name	**MOVING_Data_WP2_3_hetrec2011_lastfm**
Dataset description	This dataset contains social networking, tagging, and music artist listening information from a set of 2K users from Last.fm online music system[7]. The dataset is released in the framework of the 2nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec 2011) at the 5th ACM Conference on Recommender Systems (RecSys 2011).
Standards and metadata	Users, artists, tags and listening are stored in DAT format.
Data sharing	The dataset is publicly available for non-commercial use[8].

[6] https://grouplens.org/datasets/movielens/1m/

[7] http://www.last.fm

[8] https://grouplens.org/datasets/hetrec-2011/

Archiving and preservation	The dataset and analysis results are stored persistently at ZBW premises on protected servers. The servers are protected by established security measures for preventing unauthorised access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups. The archiving and preservation of this dataset are performed by the GroupLens Research Project at the University of Minnesota; MOVING has no involvement in this process.

Dataset name	**MOVING_Data_WP2_4_LibraryThing**
Dataset description	This dataset contains 2,056,487 million ratings of 37,232 books from 7,279 users crawled from LibraryThing[9].
Standards and metadata	Users, artists, ratings are stored in CSV format.
Data sharing	The dataset was originally publicly available, but it is no more. The sharing of the data is controlled by ZBW. The dataset is available only to the participating partners of MOVING due to possible copyright restrictions that some collected data might have. Furthermore, the dataset can be requested by other research institutions via individual agreements with ZBW.
Archiving and preservation	The dataset and analysis results are stored persistently at ZBW on protected servers. The servers are protected by established security measures for preventing unauthorised access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups.

Dataset name	***MOVING_Data_WP2_5_Aminer_DBLP***
Dataset	The DLBP Citation Network dataset (version 10)[10] includes 25,166,994 citations. The

[9] https://www.librarything.com/

[10] https://aminer.org/citation

description	dataset comprises 3,079,007 distinct citing documents published between 1936, and 2018 and 1,985,921 distinct cited documents.
Standards and metadata	The following metadata are available for each document: identifier, title, authors, venue, year, citations, references, and abstract. They are stored in JSON format.
Data sharing	The dataset is publicly available.
Archiving and preservation	The dataset and analysis results are stored persistently at ZBW premises on protected servers. The servers are protected by established security measures for preventing unauthorised access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups.

Dataset name	**MOVING_Data_WP2_6_Aminer_ACM**
Dataset description	The ACM Citation Network dataset (version 9)[11] contains 11,344,141 citations. The dataset comprises 2,385,066 distinct citing documents published between 1936, and 2016 and 2,631,128 distinct cited documents.
Standards and metadata	The following metadata are available for each document: identifier, title, authors, venue, year, citations, references, and abstract. They are stored in text format.
Data sharing	The dataset is publicly available.
Archiving and preservation	The dataset and analysis results are stored persistently at ZBW on protected servers. The servers are protected by established security measures for preventing unauthorised access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups.

Dataset name	**MOVING_Data_WP2_7_MPD**

[11] https://aminer.org/citation

Dataset description	The Million Playlist Dataset[12] contains 1,000,000 playlists created by users on the Spotify platform between January 1, 2010 and December 1, 2017.
Standards and metadata	Every playlist has a title and includes at least three unique artists and two unique albums, has a minimum number of followers and listeners and has at least 5 tracks and no more than 250 tracks. Each playlist is characterized by the playlist identifier, the playlist title, the total number of tracks in the playlist and the number of tracks included in the playlist. For each track in the playlist, the following information is included: the position in the playlist, the title, the Spotify URI, the name and Spotify URI of the primary artist, the title and the Spotify URI of the album, the duration. They are stored in JSON format.
Data sharing	The dataset was available only to the participants to the ACM RecSys challenge 2018[13] and was used by ZBW for that purpose only.
Archiving and preservation	The analysis results are stored persistently at ZBW premises on protected servers. The servers are protected by established security measures for preventing unauthorised access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups.

Dataset name	**MOVING_Data_WP2_8_MOOC_Evaluation_1**
Dataset description	This dataset contains anonymised online survey data gathered for evaluation purposes from participants of the 1st round of the MOOC[14]. Data about the following aspects on the evaluation can be found in the data set: • Demographics. • Motivation, expectation and satisfaction with the overall course and each course week.
Standards and	The data consists of eight answered questionnaires and is in .csv format.

[12] http://recsys-challenge.spotify.com/dataset

[13] http://www.recsyschallenge.com/2018/

[14] https://moving.mz.tu-dresden.de/mooc

metadata	
Data sharing	Access to the dataset is restricted to the MOVING project partners to ensure privacy protection and no abuse of the data. Results of the empirical analysis are accessible via deliverable D2.3 and will be made accessible to third parties via scientific publications, posters and project reports.
Archiving and preservation	The dataset is permanently stored on file servers of TUD. File servers are protected by applying the commonly used security measures for preventing unauthorised access and ensuring that security software is up-to-date with the latest released security patches. Preservation is ensured by regularly backups of the file systems. Backups are either conducted internally with the local administrator of the Media Center or with the help of a service partner within TUD, that offers free of charge backups with IBM Tivoli Storage System (TSM) including a guaranteed maximum 10 year time span to preserve this data. Additionally duration may be discussed with said partner.

Dataset name	**MOVING_Data_WP2_9_MOOC_Evaluation_2**
Dataset description	This dataset contains anonymised online survey data gathered for evaluation purposes from participants of the 2nd round of the MOOC. Data about the following aspects on the evaluation can be found in the data set: • Demographics. • Motivation, expectation and satisfaction with the overall course and each course week.
Standards and metadata	The dataset consists of 40 answered questionnaires and is in .csv format.
Data sharing	Access to the dataset is restricted to the MOVING project partners to ensure privacy protection and no abuse of the data. Results of the empirical analysis are made accessible to third parties via scientific publications, posters, project reports etc.
Archiving and preservation	The dataset is permanently stored on file servers of TUD. File servers are protected by applying the commonly used security measures for preventing unauthorised access and ensuring that security software is up-to-date

	with the latest released security patches. Preservation is ensured by regularly backups of the file systems. Backups are either conducted internally with the local administrator of the Media Center or with the help of a service partner within TUD, that offers free of charge backups with IBM Tivoli Storage System (TSM) including a guaranteed maximum 10 year time span to preserve this data. Additionally duration may be discussed with said partner.

Dataset name	**MOVING_Data_WP2_10_MOOC1_ForumEntries**
Dataset description	This dataset contains anonymised data gathered from the forum entries of participants of both rounds of the MOOC "Science 2.0 and open research" which was held on the MOVING platform. The data set gives insights about the following topics: • Individual presentation of the participants with their background and their expectations of the course. • Personal and professional experiences of participants with Open Science and open research methods. • Discussions on benefits and challenges of Open Science- user-generated content (e.g. embedded media like Twitter posts, presentation slides, articles, links). • Discussions on the use of social media in academia. • Discussions on the feasibility of open science workflows in different disciplines.
Standards and metadata	The dataset consists of 531 forum entries. The dataset is in .csv format.
Data sharing	Access to the dataset is restricted to the MOVING project partners to ensure privacy protection and no abuse of the data. Results of the empirical analysis are made accessible to third parties via scientific publications, posters, project reports etc.
Archiving and preservation	The dataset is permanently stored on file servers of TUD. File servers are protected by applying the commonly used security measures for preventing unauthorised access and ensuring that security software is up-to-date with the latest released security patches. Preservation is ensured by regularly backups of the file systems. Backups are either conducted internally with the local

	administrator of the Media Center or with the help of a service partner within TUD, that offers free of charge backups with IBM Tivoli Storage System (TSM) including a guaranteed maximum 10-year time span to preserve this data. Additionally duration may be discussed with said partner.

2.3 WP3 Datasets

Dataset name	**MOVING_Data_WP3_1_TRECVID**
Dataset description	This dataset is provided by NIST to the participants of the TRECVID SIN and MED tasks. It is used for developing technologies for video annotation with visual concept labels. The dataset is divided in two main parts. The first part consists of approx. 18,500 videos (354 GB, 1,400 hours) under a Creative Commons (CC) license, in MPEG-4/H.264 format, and it is typically partitioned into training (approx. 11,200 videos, 10 seconds to 6,4 minutes long; 210 GB, 800 hours total) and testing set (approx. 7300 videos, 10 seconds to 4,1 minutes long; 144 GB, 600 hours total) for video concept detection methods. The total number of concepts is 346, and the annotation of each of these videos is based on a pair of XML and TXT files; the XML file contains information about the shot segments of the video and the TXT file includes the shot-level concept-based annotation of the video via a number of positive and negative concept labels. Finally, a TXT file with metadata describing sets of relations between these concepts in the form of "concept A implies concept B" and "concept A excludes concept B", is also available. The second part is a collection of approx. 63,000 videos (736 GB, 2,520 hours) in MPEG-4/H.264 format, created by the Linguistic Data Consortium and NIST. It is used for the development of video event detection techniques and is divided in three subsets: (a) a training set with 3,000 (50 GB, 80 hours) positive or near-miss videos, and 5,000 (51 GB, 200 hours) background (i.e., negative) videos, (b) a validation set of 23,000 videos (272 GB, 960 hours), and (c) an evaluation set of 32,000 videos (363 GB, 1280 hours). The number of considered events is 20, and the ground truth for this collection is stored in CSV files. These files provide the event-based annotations of the videos by defining the list of positive or near-miss videos for each visual event.
Standards and metadata	The videos of this static dataset are in MPEG-4/H.264 format, while their annotations and metadata are in TXT, XML and CSV files. The generated results after

	processing this dataset (extracted features, if any; automatic annotation results) can be stored in XML, JSON, MPEG-7 or txt formats. They are accompanied by a document (a word file) containing metadata with sufficient information to: (a) link it to the research publications/outputs, (b) identify the founder and research discipline, and (c) appropriate key words to help users to locate the data.
Data sharing	This is a dataset created and provided to us by NIST, under specific conditions that are linked with the TRECVID benchmarking activity. Sharing of the dataset is regulated by NIST, and we comply with their requirements. We are not allowed to further share this dataset with 3rd parties. The results of our processing of the dataset are uploaded into the project's Zenodo profile.
Archiving and preservation	As stated above, a set of processing outcomes of the dataset is made available into the project's Zenodo profile. The original dataset and the analysis results are stored on the file servers of CERTH (protected by applying the commonly used security measures for preventing unauthorised access and ensuring that security software is up-to-date with the latest released security patches) and backup provisions have been made.

Dataset name	**MOVING_Data_WP3_2_ImageNET**
Dataset description	This dataset contains images of the online ImageNet collection, which is organised and managed by the Stanford and Princeton Universities. It is used for building and training Deep Convolutional Neural Networks (DCNNs) for video concept detection. In particular, ImageNet is an image dataset organised according to the WordNet hierarchy (currently only the nouns); for each node of the hierarchy, related images (often several hundreds or thousands of them) are provided. The current dataset is the one released in fall 2011 and is an updated version of the initial collection. It contains approx. 15 million images in high resolution JPEG format, which are clustered in categories that correspond to 22,000 distinct concepts of the WordNet structure. Images of each concept are quality-controlled and human-annotated.
Standards and metadata	This static dataset is composed by images that are mainly in high resolution JPEG format. The created metadata after analysing these images can be: (a) local features extracted from these images, that are stored in BIN of TXT files, and (b) the output of the trained DCNNs (i.e., the classification decision), which is stored in TXT files. These data are accompanied by a document (a word file) containing metadata with sufficient information to: (a) link it to the research publications/outputs, (b) identify

	the funder and discipline of the research, and (c) appropriate key words to help internal users to locate the data.
Data sharing	The ImageNet dataset is freely available for non-commercial research and/or educational use, by following the procedure and adopting the terms of use that are described in the ImageNet website.
Archiving and preservation	The original dataset and the results of processing it are stored on the file servers of CERTH (protected by applying the commonly used security measures for preventing unauthorised access and ensuring that security software is up-to-date with the latest released security patches) and backup provisions have been made. The archiving and preservation of this dataset are performed by the Stanford and Princeton Universities; MOVING will have no involvement in this process.

Dataset name	**MOVING_Data_WP3_3_SocialMediaWebRetrieval**
Dataset description	This dataset comprises of the web content that is collected by the MOVING platform as part of Task 3.2. Content collection is seeded by a user-defined set of topics. The dataset includes both text and multimedia content and collects from different social networks such as Twitter, Google+, Youtube etc., and from selected webpages.
Standards and metadata	The collected dataset from social networks and webpages is stored in the elasticsearch[15] database and follows the common data model format.
Data sharing	Due to the limitations imposed by Twitter (republishing a tweet is not allowed, only linking[16] or embedding[17] it is permitted) and the other social media platforms to be accessed, access to the dataset is restricted to MOVING partners. Fully anonymised aggregated analysis results are included in public reports, deliverables, and scientific publications.
Archiving and preservation	The dataset is stored persistently (i.e. guaranteed until project's end and planned to be kept long after the end of it) on a TUD server (protected by applying the commonly used security measures for preventing unauthorised access and ensuring

[15] https://www.elastic.co/products/elasticsearch

[16] https://support.twitter.com/articles/80586#

[17] https://support.twitter.com/articles/20169559#

	that security software is up-to-date with the latest released security patches) except for the multimedia content (videos), which are stored on a CERTH server for a short period of time to be processed and then are deleted. Only video metadata, which includes the processing results, are stored permanently on the TUD server. Preservation is ensured by backup of the original databases or file systems.

Dataset name	**MOVING_Data_WP3_4_GESISPublicationMetaData**
Dataset description	This dataset comprises about 3,9 million open metadata records of social science publications in German speaking countries collected at GESIS from different sources. The dataset has been made publicly available for searching via the Sowiport portal (www.sowiport.de) provided by GESIS until end of 2017. Sowiport is offline since January 2018. Until end of 2017 the dataset was updated periodically by GESIS. The dataset is useful for the TUD use case. To the best of our knowledge, there is no similar collection for German speaking countries.
Standards and metadata	Each record contains common metadata on scholarly publications such as author(s), title, abstract, keywords, classification, and bibliographical data. All records are indexed using controlled social science vocabulary. Cooperation partners uploaded data in custom format via ftp. Files were automatically parsed w.r.t. the partner-specific formatting, fed into the database and were made available on the platform. In case of errors, manual correction was done at GESIS where it is possible. There was no versioning. Entries could be overwritten but no old versions were stored, nor was the number of changes recorded.
Data sharing	Due to agreements with cooperation partners the dataset can be made available to third parties under the express conditions that the data is used solely for research purposes and may not be copied and re-used for any other purpose. Bulk downloads can be initiated for the purposes of the project using the OAI API at *sowiport.gesis.org/OAI/Home*.
Archiving and preservation	The dataset is still stored at a GESIS Vufind server in XML format. The server is protected by applying the commonly used security measures for preventing

	unauthorised access and ensuring that security software is up-to-date with the latest released security patches. Preservation is ensured by regularly backups of the original databases or file systems. Description of the sources and statistics can be viewed on the website: *sowiport.gesis.org/Database*. The full database schema can be handed out on upon request. The GESIS database SOLIS offered under Sowiport is no longer maintained. SOLIS is available at GENIOS - WISO-Net and was made available for free download at https://git.gesis.org/open-data/solis-sofis in the common data formats (rdf, json, bibtex and tsv). The other collections previously offered under Sowiport are also available elsewhere on the web (for more information see http://sowiport.gesis.org/).

Dataset name	**MOVING_Data_WP3_5_GESISOAFulltexts**
Dataset description	This dataset comprises about 9,200 open access full texts of social science publications. The full texts are available via the SSOAR repository (www.ssoar.info) provided by GESIS. The full-text server SSOAR, which is maintained at GESIS – Leibniz Institute for the Social Sciences, collects and archives literature of relevance to the social sciences and makes it available in open access on the Internet in accordance with the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. SSOAR primarily pursues the so-called “Green Road to Open Access” (OA) and sees itself as a secondary publisher of quality-controlled literature. SSOAR currently comprise about 36,000 full texts. Most of them underlie a deposit licence. Of these 5,400 full texts are openly available. The dataset is permanently updated by contributions of the social scientists. The dataset is useful for the TUD use case. To the best of our knowledge, there is no similar collection for German speaking countries.
Standards and metadata	All full texts are described by a set of common metadata, such as author(s), title, abstract, keywords, classification, and bibliographical data, and are indexed using controlled social science vocabulary. Scientists upload pdf and metadata or only metadata manually. Publication is checked manually at GESIS for appropriateness. In case of errors, manual correction

	is done at GESIS. Data is made available at a point in time specified by the author. There are no automatic means of uploading data. All entries have a version number. All versions are available on the portal.
Data sharing	Due to license issues, the dataset can be made available to third parties under the express conditions that the data is used solely for research purposes and may not be copied and re-used for any other purpose. Bulk downloads can be initiated by anyone using the OAI API at *www.ssoar.info/OAIHandler*
Archiving and preservation	The dataset is persistently stored at a GESIS DSpace server. The server is protected by applying the commonly used security measures for preventing unauthorised access and ensuring that security software is up-to-date with the latest released security patches. Preservation is ensured by regularly backups of the original databases or file systems. Some general information is available on the website: www.ssoar.info/home/veroeffentlichen-auf-ssoar.html. The schema is not described on publicly available sites, but it can be derived easily when looking at the fields offered on the portal or in the OAI xml files.

Dataset name	**MOVING_Data_WP3_6_GESISProjectMetaData**
Dataset description	This dataset comprises about 54,000 metadata record on social science research projects (SOFIS) collected by GESIS. The dataset has been made publicly available for searching via the SOFISwiki portal (https://sofis.gesis.org/sofiswiki/Hauptseite) provided by GESIS until end of 2017. SOFISwiki is offline since January 2018. The dataset was permanently updated by contributions of the social scientists. The dataset is useful for the TUD use case. To the best of our knowledge, there is no similar collection for German speaking countries. As regards DFG projects there is an overlap with the DFG project database.
Standards and metadata	Each record contains common metadata such as researcher(s), title, abstract, keywords, classification, and a number of metadata on research methods and publications. All records are indexed using controlled social science vocabulary. Scientists uploaded or modified metadata of their project entries manually. All edits

	have been checked manually at GESIS for appropriateness. In case of errors, manual correction were done at GESIS. Additionally, bulk imports from cooperation partners were done regularly. Projects that have been provided by DFG were not being processed due to overlaps with the existing database. There was no decided versioning, but the display format of Semantic Mediawiki allowed viewing of previous versions for single projects.
Data sharing	Due to the terms of use of the SOFISwiki (Creative Commons License CC BY-NC-SA 3.0 DE), the dataset can be made available to cooperation partners of GESIS within the framework of common projects under the express conditions that the data is used solely for research purposes and may not be copied and re-used for any other purpose. There is no repository for sharing the whole dataset, but it can be provided by GESIS upon request for the purposes of the MOVING project.
Archiving and preservation	The dataset has been stored at a GESIS Semantic MediaWiki server. The server was protected by applying the commonly used security measures for preventing unauthorised access and ensuring that security software is up-to-date with the latest released security patches. Preservation was ensured by regularly backups of the original databases or file systems. The GESIS database SOFIS is no longer maintained. SOFIS is available at GENIOS - WISO-Net and was made available for free download at https://git.gesis.org/open-data/solis-sofis in the common data formats (rdf, json, bibtex and tsv).

Dataset name	**MOVING_Data_WP3_7_InteractionData**
Dataset description	This dataset contains the data generated through the use of the MOVING platform. Consequently it consists of interaction events at a different level of granularity: • Low-level events are those triggered through the use of the keyboard, mouse, touchscreen or mouse wheel. • Higher-level events are semantically more meaningful and describe specific actions on the platform: “search for X”, “open Y” and similar. A row of this dataset looks like: *user, URL, timestamp, event, object of the event*. This dataset is key to accomplish Task 3.3 in order to infer knowledge acquisition from user behavior on the MOVING platform.

Standards and metadata	This dataset is formatted as CSV files.
Data sharing	While this data is open for the members of the MOVING consortium, it cannot be shared, as it was generated in TUD servers and due to German Data Protection Law, data should remain there. Data have been analysed by remotely accessing TUD servers and they have never been downloaded elsewhere.
Archiving and preservation	Due to the above reasons the data are kept in TUD servers. The file servers themselves are protected by applying the commonly used security measures for preventing unauthorised access and ensuring that security software is up-to-date with the latest released security patches. Preservation is ensured by regularly backups of the file systems. These backups are conducted with the help of a service partner within TUD, that offers free of charge backups with IBM Tivoli Storage System (TSM) including a guaranteed maximum 10 year time span to preserve this data. Additionally duration may be discussed with said partner.

Dataset name	**MOVING_Data_WP3_8_BTC2014**
Dataset description	The Billion Triples Challenge 2014 (BTC)[18] dataset contains structured data from various domains like Government, Publications, Life sciences, User-generated content, Cross-domain, Media, Geographic and Social Web. The latest dataset from 2014 covers 47,560 different pay-level domains and contains 4,090,758,596 (4 Billion) RDF-quads in total. All quads are available in one dump with a volume of 1.1TB (unzipped) or 52GB (compressed). Besides the raw data, the BTC dataset comes with additional metadata information regarding origin and statistical information of the sources. The dataset is crawled from the public web via the open source program LDSpider[19]. Since the data is crawled from the Web, it is of varying quality regarding structure and content. The BTC2014 dataset is relevant for researchers analysing Linked Open Data (LOD) and developing LOD-based applications.
Standards and	All data including the metadata is formatted in standards defined by the W3C,

[18] http://km.aifb.kit.edu/projects/btc-2014/

[19] https://github.com/ldspider/ldspider

metadata	namely the RDF format and N-Quads.
Data sharing	The original BTC 2014 dataset is hosted by KIT in Karlsruhe and can be freely downloaded. The procedure is described on the website of the dataset.
Archiving and preservation	The dataset is stored persistently at ZBW on dedicated servers. The servers are protected by established security measures for preventing unauthorised access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups. The dataset is also partly integrated in the MOVING platform. Specifically, only bibliographic metadata have been included.

Dataset name	**MOVING_Data_WP3_9_LODLaundromat**
Dataset description	The LOD Laundromat[20] provides access to a very large collection of Linked Open Data (LOD). The creators do not claim to provide a "new dataset, but rather a uniform point of entry to a collection of cleaned siblings of existing datasets." (Beek et al., 2014). The latest dump from 2016 contains 38,606,408,854 (38 Billion) RDF-quads in total. All quads are available in one dump with a volume of 291GB (gzipped N-Triples). The dataset is crawled from the public web via the open source program WashingMachine[21]. Since the data is crawled from the Web, it is of varying quality regarding structure and content. Although the data is cleansed, it does not mean that any kind of professional content review has been performed. The LODLaundromat dataset is relevant for researchers analysing Linked Open Data (LOD) and developing LOD-based applications.
Standards and metadata	All data including the metadata is formatted in standards defined by the W3C, namely the RDF format and N-Quads.
Data sharing	The original LOD Laundromat dataset can be accessed via the Wardrobe[22]. The procedure is described on the website of the dataset.

[20] http://lodlaundromat.org/

[21] https://github.com/LOD-Laundromat/LOD-Laundromat

[22] http://lodlaundromat.org/wardrobe/

Archiving and preservation	The dataset is stored persistently at ZBW on dedicated servers. The servers are protected by established security measures for preventing unauthorised access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups.

Dataset name	**MOVING_Data_WP3_10_DyLDO**
Dataset description	The Dynamic Linked Data Observatory (DyLDO)[23] dataset contains weekly snapshots of larger crawls from the Linked Open Data (LOD) cloud starting with 06.05.2012. On average (over the first 29 datasets) there are 1,738.6 pay-level domains in the dataset with a total of 94,725,595 quads bibliography (Käfer et al., 2013). The weekly crawls are stored in separate dumps. As in the BTC2014 dataset, the DyLDO dataset also contains structured data from various domains. Although the snapshots are smaller compared to the BTC2014, they are of added value since they capture the evolution of the LOD cloud over time. Furthermore, also the Dynamic Linked Data Observatory uses the LDSpider to conduct the weekly crawls. The DyLDO dataset is relevant for researchers analysing the temporal evolution of Linked Open Data (LOD) and developing LOD-based applications considering the time aspect.
Standards and metadata	All data is formatted in standards defined by the W3C, namely the RDF format and N-Quads.
Data sharing	The dataset is provided by Karlsruhe Institute of Technology in Germany. The weekly crawls can be freely downloaded. The procedure is described on the website of the dataset.
Archiving and preservation	The dataset is stored persistently at ZBW on dedicated servers. The servers are protected by established security measures for preventing unauthorised access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups.

Dataset name	**MOVING_Data_WP3_11_ZBWEconomicsDataset**

[23] http://km.aifb.kit.edu/projects/dyldo/

Dataset description	The ZBWEconomicsDataset is provided by the partner institution ZBW – Leibniz Information Centre for Economics. ZBW is running a search portal, called EconBiz[24], for economics' scientific publications. From EconBiz, ZBW obtained 1 million URLs of open access scientific publications and generated a dataset of about 413,000 full-texts in PDF format and the metadata (e.g., authors, title, year of publication). From the 413,000 scientific publications in the economics domain, 280,000 publications are in English. The remaining publications cover 41 other languages, most notably German, French and Spanish. In addition to the PDF format, the publications are also converted to plain ASCII text format (TXT). Furthermore, from the English publications we extracted about 200,000 images which are most likely scholarly figures. From these figures, we randomly selected and manually annotated 121 scholarly figures. The figures vary in type (e.g., bar chart, pie chart, map), quality and topic. Ground truth information containing the position, orientation for each text line is stored in TSV and JSON format. Information about the origin of the individual figures is provided in textual form in form of an identifier to the EconBiz portal. The scientific publications of the ZBWEconomicsDataset are relevant for researchers working on information retrieval, recommendation, and document classifications tasks. The annotated set of figures is relevant for researchers who try to improve the indexing of figure-like images based on the text inside these figures.
Standards and metadata	The full-texts are available in plain ASCII format (TXT) and PDF format. The metadata are provided in JSON format. The ground truth information for the manually annotated scholarly figures is available in TSV and JSON format.
Data sharing	The sharing of the data is controlled by ZBW. The dataset is available only to the participating partners of MOVING due to possible copyright restrictions that some collected publications might have. Furthermore, the dataset can be requested by other research institutions via individual agreements with ZBW. The corpus of the 121 manually annotated figures, which are from open access publications, is publicly available[25]
Archiving and preservation	The dataset is stored persistently on dedicated servers of ZBW. The servers are protected by established security measures for preventing unauthorised access and

[24] http://www.econbiz.de/

[25] http://www.kd.informatik.uni-kiel.de/en/research/software/text-extraction-files/econbiz-dataset

	ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups.

Dataset name	**MOVING_Data_WP3_12_NYT**
Dataset description	The New York Times dataset is organised and managed by the Linguistic Data Consortium (LDC)[26]. It contains a set of more than 1.8 million articles that have been published between 1st January 1987 and 19th of June 2007. More than 650,000 articles contain a professional summary and more than 1.5 million articles have been manually annotated by librarians with various tags for places, people, organisations and topics. The dataset is available in form of XML files. Each file contains information about a single article. For instance[27] the publication date, section, author, title and contents. The dataset is relevant for researchers working on information retrieval, recommendation, and document classifications tasks.
Standards and metadata	A copy of the NYT dataset is available in form of XML documents in the NITF format. The data is provided together with an open source java tools for parsing the documents into memory objects.
Data sharing	The dataset is provided and managed by the Linguistic Data Consortium (LDC), The Trustees of the University of Pennsylvania. The dataset is individually licensed (per organisation or individual) for some fee[28]. The data is owned by the New York Times Company.
Archiving and preservation	The dataset and analysis results are stored persistently at ZBW on protected servers. The servers are protected by established security measures for preventing unauthorised access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups. The archiving and preservation of this dataset are performed be the LDC. MOVING has no involvement in this process.

[26] https://www.ldc.upenn.edu

[27] https://catalog.ldc.upenn.edu/desc/addenda/LDC2008T19_large.jpg

[28] https://catalog.ldc.upenn.edu/LDC2008T19

Dataset name	**MOVING_Data_WP3_13_Yago**
Dataset description	YAGO[29] is a knowledge base created by the Max Planck Institute in Saarbrucken and derived from Wikipedia[30], WordNet[31] and GeoNames[32]. The YAGO-NAGA project started in 2006 with the aim to build a "*highly accurate knowledge base of common facts in a machine-processible representation*". In 2012, the second version of YAGO (YAGO2) has been released (Hoffart, et al., 2013), followed by the third version (YAGO3) in 2015. YAGO3 contains more than 10 million entities with more than 120 million multilingual properties. YAGO's accuracy has been manually evaluated using a sample of facts and the extrapolated accuracy is between 90.84% and 99.22%[33]. The dataset is relevant for researchers working on entity extraction, entity classification, document retrieval and others.
Standards and metadata	YAGO is available in RDF/Turtle and TSV format.
Data sharing	YAGO is provided by Max Planck Institute for Informatics in Saarbrucken and is licensed under the CC 3.0 license.
Archiving and preservation	The dataset and analysis results are stored persistently at ZBW on protected servers. The servers are protected by established security measures for preventing unauthorised access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups. The archiving and preservation of this dataset is done by the Max Planck Institute for Informatics in Saarbrucken together with the Databases and Information Systems Group[34] and the DBWeb team of Télécom ParisTech[35]. MOVING has no involvement in this process.

[29] www.mpi-inf.mpg.de/YAGO/

[30] https://www.wikipedia.org/

[31] https://wordnet.princeton.edu/

[32] http://www.geonames.org/

[33] http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/statistics/

[34] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/

[35] http://dbweb.enst.fr/research/

Dataset name	**MOVING_Data_WP3_14_ScientificPublicationRetrieval**
Dataset description	The dataset consists of a set of publicly available scientific publications that are obtained via established crawlers and protocols from corresponding open access archives. All publications are associated with their domain (e.g. economics, computer science). The publications are stored in PDF format as well as in plain ASCII format (TXT). The MOVING crawlers keep the documents up-to-date. The documents from the economics-related repositories (e.g. AgEcon[36] and Munich Personal RePEc[37]) Archive support the EY use case, while the crawled data from the computer science related repositories (e.g. CiteseerXData and arXiv Bulk Data[38]) support the TUD use case. The scientific publications of this dataset are relevant for researchers working on information retrieval, recommendation, and document classifications tasks. Although the original idea was to rely on the OCLC OAI Harvester2[39] toolkit, these documents have been integrated from CORE[40] because the latter provided a higher quality of metadata (more complete and consistent).
Standards and metadata	The publications are stored in PDF and plain ASCII (TXT). The metadata of the dataset are formatted and stored in JSON and CSV formats. The metadata includes the following information such as the domain, title, authors, venue and year.
Data sharing	The publications are collected from open access repositories such as those mentioned above. The dataset is available only to the participating partners of MOVING due to possible copyright restrictions that some collected publications might have.
Archiving and preservation	The dataset and the analysis results are stored persistently at ZBW on protected servers. The servers are protected by established security measures for preventing unauthorised access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups. The archival of the publications is performed by the crawled archive providers. MOVING has no involvement in this process.

[36] http://ageconsearch.umn.edu/

[37] https://mpra.ub.uni-muenchen.de/

[38] https://arxiv.org/help/bulk_data

[40] https://core.ac.uk/about

Dataset name	MOVING_Data_WP3_15_DBPedia
Dataset description	DBpedia is a Knowledge Base derived from Wikipedia. In 2007 the first version of DBpedia has been published. The project was initiated by the Free University of Berlin, the University of Leipzig[41] and OpenLink Software[42]. In the English version of the dataset, DBpedia currently has 4.22 million entities, including at least 1,445,000 people, 735,000 places, 123,000 music albums, 87,000 films and 19,000 video games that are described and licensed under a creative commons (CC) license. The dataset is stored using the Resource Description Framework (RDF) and queries can be made using SPARQL[43]. The dataset is relevant for researchers working on entity extraction, entity classification, document retrieval and others.
Standards and metadata	The DBPedia dataset is not only available in RDF, but also as a JSON and CSV tabular version. Since DBPedia extracts information from Wikipedia, The dataset provides structured information about the articles in forms of RDF Triples. For instance, the birth date of a person takes the form (person, date of birth, date).
Data sharing	The textual content is reusable in the terms of the GNU Free Documentation License (GFDL) and the Creative Commons Attribution-Share-Alike 3.0 License.
Archiving and preservation	The dataset and the analysis results are stored persistently at ZBW on protected servers. The servers are protected by established security measures for preventing unauthorised access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups. The archiving and preservation of this dataset are performed by the Free University of Berlin, the University of Leipzig and OpenLink Software. MOVING has no involvement in this process.

Dataset name	MOVING_Data_WP3_16_RCV1
Dataset	The Reuters' dataset, known as "Reuters Corpus, Volume 1" or RCV1 is provided by

[41] https://www.zv.uni-leipzig.de/en/

[42] http://www.openlinksw.com/

[43] http://wiki.dbpedia.org/OnlineAccess

description	NIST for research purposes. The RCV1 contains more than 800.000 manually-labelled Reuters' news stories. All news stories were written in English and published between 20th of August 1996 and 19th of August 1997. The news stories are organised in files. Each news article is organised in a separate XML file. Each XML file contains some information about an article. For instance title, publisher, contents and location. The news articles are categorised and controlled using the following three main vocabularies[44]: (a) Topics (126 topics), (b) Industries (870) and (c) Regions (366 geographic codes). All documents are contained in one dump with a volume of 2.5GB. The dataset is relevant for researchers working on information retrieval, recommendation, and document classifications tasks.
Standards and metadata	The dataset files have been formatted in XML.
Data sharing	The copyright of the dataset is reserved to Reuters Ltd and/or Thompson Reuters. The dataset is provided and regulated by NIST. It can be used for research purposes. Every user or organisation has to request his/its own copy[45][46].
Archiving and preservation	The dataset and analysis results are stored persistently at ZBW on protected servers. The servers are protected by established security measures for preventing unauthorised access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups. The archiving and preservation of this dataset are performed by NIST. MOVING has no involvement in this process.

Dataset name	**MOVING_Data_WP3_17_EnglishLanguageWikimedia**
Dataset description	Wikimedia Foundation, the parent organisation of Wikipedia, published a dump[47] of more than 4.4 million Wikipedia articles containing more than 1.9 billion words. Wikimedia updates the dataset regularly, roughly twice a month. A free copy of this dataset is available in form of XML and SQL files. Each file contains information

[44] http://jmlr.csail.mit.edu/papers/volume5/lewis04a/lewis04a.pdf
[45] http://trec.nist.gov/data/reuters/ind_appl_reuters_v4.html
[46] http://trec.nist.gov/data/reuters/org_appl_reuters_v4.html
[47] https://dumps.wikimedia.org/

	about an article. For instance the publication date, section, author, title and contents.[48] The dataset is relevant for researchers working on information retrieval, recommendation, and document classifications tasks.
Standards and metadata	The dataset is available in form of XML and/or SQL files.
Data sharing	All text contents can be used in the terms of the GNU Free Documentation License (GFDL) and the Creative Commons Attribution-Share-Alike 3.0 License.
Archiving and preservation	The dataset and analysis results are stored persistently at ZBW on protected servers. The servers are protected by established security measures for preventing unauthorised access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups. The archiving and preservation of this dataset are performed by the Wikimedia Foundation. Wikimedia usually updates the dataset twice a month. MOVING has no involvement in this process.

Dataset name	**MOVING_Data_WP3_18_CHIME**
Dataset description	The CHIME[49] dataset was created by the Center for Information Mining and Extraction, School of Computing, National University of Singapore (NUS). It consists of two sets of figures, one based on real world data created by Yang et al.[50] (Yang, et al. 2006) and one based on synthetic data created by Jiuzhou[51], each having different numbers of bar charts, pie charts and line graphs. The synthetic set contains 85 figures while the other subset contains 115 figures. It was published to promote research activities involving figures. Each figure, from the dataset created by NUS, is accompanied with ground truth information in plain ASCII (TXT) and XML format which contain the position and content of all text and graphic elements inside the figure, but no information about their orientation. A transformed set of ground truth information about the text

[48] https://catalog.ldc.upenn.edu/desc/addenda/LDC2008T19_large.jpg

[49] https://www.comp.nus.edu.sg/~tancl/ChartImageDataset.htm

[50] https://www.comp.nus.edu.sg/~tancl/publications/c2006/das06-062.pdf

[51] https://www.comp.nus.edu.sg/~tancl/ChartImageDatabase/Report_Zhaojiuzhou.pdf

	content in TSV and JSON format was created by ZBW and is available similar to the Economics Figures dataset. The annotated set of figures is relevant for researchers who try to improve the indexing of figure-like images based on the text inside these figures.
Standards and metadata	For representing the metadata and gold standard the TSV and JSON formats are used.
Data sharing	The dataset is publicly available.[52] The datasets with the extended ground truth information provided by ZBW are also publicly available.[53]
Archiving and preservation	The dataset and analysis results are stored persistently at ZBW on protected servers. The servers are protected by established security measures for preventing unauthorised access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups. The archiving and preservation of the figure dataset is performed by NUS. MOVING has no involvement in this process. The archiving and preservation of the extended gold standard is conducted by ZBW and the same mechanisms are applied as described above for storing and managing the dataset and analysis results.

Dataset name	**MOVING_Data_WP3_19_EconBizRecSysEvaluationResult**
Dataset description	The EconBizRecSysEvaluationDataset[54] contains the evaluation results from a previously built recommender system based on the assessments of 123 participants after analysing their Twitter profiles[55]. The evaluation results have been generated from twelve different strategies, each of which contains five scientific publication recommendations and its corresponding user assessment. The recommended scientific publications are a subset of the MOVING_Data_WP3_11_ ZBWEconomicsDataset (mentioned above).
Standards and	The dataset is available in XML and/or SQL file format.

[52] https://www.comp.nus.edu.sg/~tancl/ChartImageDataset.htm
[53] http://www.kd.informatik.uni-kiel.de/en/research/software/text-extraction
[54] https://datorium.gesis.org/xmlui/handle/10.7802/1224
[55] http://ieeexplore.ieee.org/abstract/document/7559581

metadata	
Data sharing	The dataset is publicly available with unrestricted download possibility for all users of GESIS data sharing repository, without requesting access permission.
Archiving and preservation	The dataset and analysis results are stored persistently at ZBW on protected servers. The servers are protected by established security measures for preventing unauthorised access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups. The archiving and preservation of this dataset are performed by GESIS. MOVING has no involvement in this process.

Dataset name	**MOVING_Data_WP3_20_NTCIR-2-IR**
Dataset description	The NTCIR2 dataset consists of 49 topics, 400k Japanese documents and 130k English documents. They have been extracted from the NACSIS Academic Conference Paper Database, collected between 1997, and 1999, and NACSIS grant-in-aid scientific research database, collected between 1988, and 1997. The documents are composed of a title and abstract field. The topics consist of the fields' title, description and narrative. Additionally, two sets of relevance scores are provided.
Standards and metadata	The dataset is available in form of TGZ files and PDF manuals, which contain more information about the dataset.
Data sharing	The dataset can only be used for research purposes. The dataset is distributed by National Institute of Informatics (NII) in JAPAN.
Archiving and preservation	The dataset and analysis results are stored persistently at ZBW on protected servers. The servers are protected by established security measures for preventing unauthorised access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups. The archiving and preservation of this dataset are performed by NII. MOVING has no involvement in this process.

Dataset name	**MOVING_Data_WP3_21_PubMed-2-IR**

Dataset description	PubMed consists of around 27 million citations for biomedical literature from MEDLINE, life science journals, and online books. Some of the citations include links to full-text content from PubMed Central and publisher web sites. From PubMed central we obtained a subset of 646,655 full-text open-access English articles.
Standards and metadata	The publications are stored in PDF and plain ASCII (TXT).
Data sharing	The dataset can only be used for research purposes. The dataset is distributed by U.S. National Library of Medicine (NLM)[56].
Archiving and preservation	The dataset and analysis results are stored persistently at ZBW premises on protected servers. The servers are protected by established security measures for preventing unauthorised access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups. The archiving and preservation of this dataset are performed by NLM. MOVING has no involvement in this process.

Dataset name	**MOVING_Data_WP3_22_IREON-2-IR**
Dataset description	The dataset consists of 27,575 full-text politics' publications in English. The documents covers mainly the following subjects: Foreign and security policy, International organizations and institutions, Theory of international relations, Economic and development cooperation, Country studies on, politics/economy/society, Regional studies, European policy, Transatlantic relations, Foreign cultural policy, Climate, environment and energy policy.
Standards and metadata	The publications are stored in PDF and plain ASCII (TXT).
Data sharing	The dataset can only be used for research purposes. The dataset is distributed by the German information network international Relations and Area Studies[57].

[56] https://www.nlm.nih.gov/

[57] http://www.fiv-iblk.de/eindex.htm

Archiving and preservation	The dataset and analysis results are stored persistently at ZBW premises on protected servers. The servers are protected by established security measures for preventing unauthorised access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups. The archiving and preservation of this dataset are performed by The German information network international Relations and Area Studies. MOVING has no involvement in this process.

Dataset name	**MOVING_Data_WP3_23_Wiki-page-views**
Dataset description	The dataset[58] consists of a statistics about the hourly page views for around 30 million Wikipedia articles. The database statistics covers a period between 2007 and 2017.
Standards and metadata	The dataset is available in a CSV file format.
Data sharing	The dataset can only be used for research purposes. The dataset is distributed by Wikimedia[59].
Archiving and preservation	The dataset and analysis results are stored persistently at ZBW premises on protected servers. The servers are protected by established security measures for preventing unauthorised access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups. The archiving and preservation of this dataset are performed by Wikimedia. MOVING has no involvement in this process.

Dataset name	**MOVING_Data_WP3_24_Twitter-Presidential-Election-2016**
Dataset	The dataset[60] contains around 280 million tweets' identifiers Sampled between July

[58] https://dumps.wikimedia.org/other/pagecounts-ez/

[59] https://wikitech.wikimedia.org/wiki/Analytics/Archive/Data/Pagecounts-raw

[60] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/PDI7IN

description	13 and November 10, 2016. The dataset contains 12 different collections related to the presidential elections in the United Stated in 2016. The collections represent the Tweets about certain events such as first presidential debate or republican party.
Standards and metadata	The tweets identifiers are stored in plain ASCII (TXT).
Data sharing	The dataset can only be used for research purposes. The dataset is distributed by GWU Libraries Dataverse (George Washington University).
Archiving and preservation	The dataset and analysis results are stored persistently at ZBW premises on protected servers. The servers are protected by established security measures for preventing unauthorised access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups. The archiving and preservation of this dataset are performed by GWU Libraries Dataverse. MOVING has no involvement in this process.

Dataset name	**MOVING_Data_WP3_25_ACL**
Dataset description	The ACL Anthology Network dataset is a collection of research papers of different Association of Computational Linguistics (ACL) venues. It was originally created by Mark Thomas Joseph and is currently being maintained by Yale University's LILY Group and our copy contains 22,486 full-text documents as well as citation information.
Standards and metadata	The documents are stored as TXT files while their metadata are stored in JSON format.
Data sharing	The original dataset is publicly available.[61]
Archiving and preservation	The dataset and analysis results are stored persistently at ZBW premises on protected servers. The servers are protected by established security measures for preventing unauthorised access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily

[61] http://tangra.cs.yale.edu/newaan/

	professional backups. The archiving and preservation of the original dataset is performed by Yale University's LILY Group. MOVING has no involvement in this process.

Dataset name	**MOVING_Data_WP3_26_DeGruyter**
Dataset description	The DeGruyter is a collection of ground truth information for text extraction of 120 figures from books by DeGruyter that was created at ZBW. The ground truth information contains the position and orientation for each text line and is stored in TSV and JSON format. Information about the origin of the individual figures is provided in textual form. The annotated set of figures is relevant for researchers who try to improve the indexing of figure-like images based on the text inside these figures.
Standards and metadata	For representing the gold standard TSV and JSON formats are used.
Data sharing	The dataset is publicly available.[62]
Archiving and preservation	The dataset and analysis results are stored persistently at ZBW premises on protected servers. The servers are protected by established security measures for preventing unauthorised access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups.

Dataset name	**MOVING_Data_WP3_27_TimBL**
Dataset description	The TimBL dataset contains about 11 million quads. The crawl was conducted in 2011 with a breadth-first search starting from the FOAF profile of Tim Berners-Lee. In contrast to the DyLDO datasets and the BTC2014 dataset, the TimBL dataset was not crawled starting with a distributed set of seed URIs. Thus, it has a different characteristic partially comparable to a depth-first crawled dataset. It is used in

[62] http://www.kd.informatik.uni-kiel.de/en/research/software/text-extraction

	previous experiments by Konrath et al., 2012 (Konrath et al., 2012).
Standards and metadata	All data is formatted in standards defined by the W3C, namely the RDF format and N-Quads.
Data sharing	The dataset is stored at ZBW. It can be shared amongst the consortium upon request.
Archiving and preservation	The dataset is stored persistently at ZBW premises on dedicated servers. The servers are protected by established security measures for preventing unauthorised access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups.

Dataset name	**MOVING_Data_WP3_28_BSBM**
Dataset description	The Berlin SPARQL Benchmark[63] (BSBM) defines a suite of benchmarks for comparing the performance of systems using RDF and SPARQL across architectures. The Berlin SPARQL Benchmark (BSBM) consists of: (a) benchmark dataset, which is scalable to different sizes based on a scale factor, (b) three query mixes measuring the performance of RDF stores against the requirements of different common use cases, and (c) a data generator and a test driver for the benchmark.
Standards and metadata	All data is formatted in standards defined by the W3C.
Data sharing	The dataset is publicly available.
Archiving and preservation	The dataset is stored persistently at ZBW premises on dedicated servers. The servers are protected by established security measures for preventing unauthorised access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups.

Dataset name	**MOVING_Data_WP3_29_LUBM**

[63] http://wifo5-03.informatik.uni-mannheim.de/bizer/berlinsparqlbenchmark/spec/index.html

Dataset description	The Lehigh University Benchmark[64] (LUBM) is developed to facilitate the evaluation of Semantic Web repositories in a standard and systematic way. The benchmark is intended to evaluate the performance of those repositories with respect to extensional queries over a large data set that commits to a single realistic ontology. The dataset consists of a university ontology, customisable and repeatable synthetic data, a set of test queries, and several performance metrics.
Standards and metadata	All data is formatted in standards defined by the W3C.
Data sharing	The dataset is publicly available.
Archiving and preservation	The dataset is stored persistently at ZBW premises on dedicated servers. The servers are protected by established security measures for preventing unauthorised access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups.

Dataset name	**MOVING_Data_WP3_30_Wikidata[65]**
Dataset description	Wikidata[66] is a free and open knowledge base that can be read and edited by both humans and machines. It is hosted by the Wikimedia Foundation[67], and its first version was published in 2012. It is the central storage for the structured data of the Wikimedia project including Wikipedia, Wikivoyage, Wikisource, etc. It is a document-oriented database, and its basic unit is the item, which represents a topic (or an administrative page to maintain Wikipedia) and is uniquely identified. Wikidata accounts for 46,013,480 items in March 2018. RDF exports are also available[68] and queries can be made using SPARQL[69]. The dataset is relevant for researchers working on entity extraction, entity classification, document retrieval

[64] http://swat.cse.lehigh.edu/projects/lubm/

[65] The use of this dataset is planned also for WP2 in the context of semantic profiling and the recommender system.

[66] http://www.wikidata.org/

[67] https://en.wikipedia.org/wiki/Wikimedia_Foundation

[68] http://tools.wmflabs.org/wikidata-exports/rdf/exports.html

[69] http://wiki.dbpedia.org/OnlineAccess

	and other text and data mining tasks.
Standards and metadata	As previously mentioned, Wikidata is based on items. An item can have one or more statements. A statement is a key-value pair that consists of a property (the key) and a value linked to the property. The dataset dumps are available in JSON, RDF, and XML[70] and results from SPARQL queries can be downloaded in JSON, TSV and CSV format.
Data sharing	The data in Wikidata is published under the Creative Commons Public Domain Dedication 1.0[71], which allows copying, modifying, distributing and performing the data, even for commercial purposes, without asking for permission.
Archiving and preservation	The dataset and the analysis results are stored persistently at ZBW premises on protected servers. The servers are protected by established security measures for preventing unauthorised access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups. The archiving and preservation of this dataset are performed by the Wikimedia Foundation. MOVING has no involvement in this process.

Dataset name	**MOVING_Data_WP3_31_CORE_DB**
Dataset description	CORE[72] harvests openly accessible content freely availability on the public internet. The CORE DB comprises metadata records as well as full-texts of scientific literature covering various domains. The latest dump from March 2018 aggregates literature from 3,753 repositories[73] and includes 123M metadata records, 85.6M records with abstract, and 9.8M records with full-text.
Standards and metadata	All data is formatted in JSON in a single data schema.
Data sharing	The dataset is publicly available. We obtained the rights by the CORE team to enrich

[70] https://www.wikidata.org/wiki/Wikidata:Database_download

[71] https://creativecommons.org/publicdomain/zero/1.0/

[72] https://core.ac.uk/about

[73] https://core.ac.uk/repositories

	our data corpus using the CORE DB.
Archiving and preservation	The dataset is stored persistently at ZBW premises on dedicated servers. The servers are protected by established security measures for preventing unauthorised access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups. It is also partly integrated in the MOVING platform. Some documents were excluded for example because they were duplicates of documents integrated from other sources or because they did not comply with the common data model (see deliverable D3.1, D3.2 and D3.3 for the common data model details).

Dataset name	**MOVING_Data_WP3_32_ZBWECONISMetadataDataset**
Dataset description	The ZBWECONISMetadataDataset is provided by the partner institution ZBW – Leibniz Information Centre for Economics. The dataset contains the online catalogue ECONIS. It consists of 4.7 million meta data records for printed and electronic economics literature from all over the world.
Standards and metadata	The metadata are provided in JSON format.
Data sharing	The sharing of the data is controlled by ZBW. The dataset is available only to the participating partners of MOVING due to possible copyright restrictions that some collected publications might have. Furthermore, the dataset can be requested by other research institutions via individual agreements with ZBW.
Archiving and preservation	The dataset is stored persistently on dedicated servers of ZBW. The servers are protected by established security measures for preventing unauthorised access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups.

Dataset name	**MOVING_Data_WP3_33_LawAndRegulationsDataset**

Dataset description	The LawAndRegulationsDataset is provided by Wolters Kluwer (WK)[74]. The dataset contains a package of German and European laws. Furthermore, the dataset contains detailed information about these laws. For example, the evolution of certain articles, the announcement and the commencement dates, and the legislators of these laws.
Standards and metadata	The metadata are provided in XML format.
Data sharing	The sharing of the data is controlled by Wolters Kluwer. We obtained the rights from WK to use this dataset.
Archiving and preservation	The dataset is stored persistently on dedicated servers of ZBW. The servers are protected by established security measures for preventing unauthorised access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups.

Dataset name	**MOVING_Data_WP3_34_LectureVideoFragmentationDataset**
Dataset description	The dataset is a large-scale lecture video dataset consisting of artificially-generated lectures, and the corresponding ground-truth fragmentation. The dataset deals with the lack of proper lecture video fragmentation datasets and it is used for the purpose of evaluating lecture video fragmentation techniques.
Standards and metadata	The dataset is created following well-known approaches from the document fragmentation field. For creating the dataset, 1498 speech transcript files (generated automatically by ASR software) were used from the world's biggest academic online video repository, the VideoLectures.NET[75]. These transcripts correspond to lectures from various fields of science, such as Computer science, Mathematics, Medicine, Politics etc. In order to create the synthetic video lectures, all transcripts were randomly split in fragments, the duration of which ranges between 4 and 8 minutes. Each synthetic lecture was then assembled by combining (stitching) exactly 20

[74] https://wolterskluwer.com

[75] http://videolectures.net/

	randomly selected fragments. 300 such artificially-generated lectures are included in the released dataset. Each such lecture file has a mean duration of about 120 minutes, thus the dataset contains altogether about 600 hours of artificially-generated lectures. Every pair of consecutive fragments in these lectures originally comes from different videos; consequently the point in time where such two fragments are joined is a known ground-truth fragment boundary. All these boundaries form the dataset's ground truth. We should stress that we do not generate the corresponding video files for the artificially-generated lectures (only the transcripts), and one should not try to reverse-engineer the dataset creation process so as to use in some way the visual modality for detecting the fragments in this dataset. The 300 artificially-generated transcripts and the corresponding ground-truth fragmentations are stored in TXT files.
Data sharing	The artificially-generated lecture video fragmentation dataset is provided for academic, non-commercial use only.
Archiving and preservation	The dataset is freely available into the project's Zenodo profile. The original lectures' transcripts and the analysis results are stored on the file servers of CERTH (protected by applying the commonly used security measures for preventing unauthorised access and ensuring that security software is up-to-date with the latest released security patches) and backup provisions have been made.

Dataset name	**MOVING_Data_WP3_35_NeLO**
Dataset description	This dataset is a collection of various statistics for vocabularies in the Linked Open Data Cloud[76] which have more than one version. We considered the available versions of the vocabularies in a period of time of over 17 years from the Linked Open Vocabulary repository[77]. Reusing terms results in a Network of Linked vOcabularies (NeLO), where the nodes are the vocabularies that use at least one term from some other vocabulary and thus depend on each other. We analyzed static parameters of NeLO such as its size, density, average degree, and the most important vocabularies at certain points in time and we include data on NeLO changes over time. Specifically, we measured the impact of a change in one vocabulary to others, how the reuse of terms

[76] https://lod-cloud.net/

[77] http://lov.okfn.org/dataset/lov/

	changes, and the importance of vocabularies changes.
Standards and metadata	The data are available in CSV, RDF and JSON format.
Data sharing	The dataset is publicly available.[78]
Archiving and preservation	The dataset is stored persistently on dedicated servers of ZBW. The servers are protected by established security measures for preventing unauthorised access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups.

2.4 WP5 Datasets

Dataset name	**MOVING_Data_WP5_1_MOVINGPublications**
Dataset description	This dataset contains manuscripts reporting the conducted scientific work in MOVING, which have been accepted for publication in high-quality peer-reviewed journals and conferences. All these publications include a statement with acknowledgement to the MOVING project, while their content vary from the description of specific analysis techniques, to established evaluation datasets and individual components and parts of the MOVING platform.
Standards and metadata	Most commonly, these documents were stored in PDF format. Each document was also accompanied by: (a) a short description with the abstract of the publications, (b) the LaTeX-related BIB file with its citation, and (c) details about the venue (e.g. conference, workshop or benchmarking activity) or journal where it was published. This dataset has been extended whenever new submitted works were accepted for publication in conferences or journals. A simple log file of the performed updates of the dataset is maintained by CERTH in the project wiki (hosted by a CERTH server).
Data sharing	This dataset is publicly available, following the guidelines of the EC[79] for open access to scientific publications and research data in Horizon2020.

[78] https://sites.google.com/view/nelo-evolution/statistics

[79] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

Archiving and preservation	Self-archiving (also known as "green open access") was applied for ensuring open access to these publications. According to this archiving policy the author(s) of the publication archived (deposit) the published article or the final peer-reviewed manuscript in online repositories: (a) personal webpage(s), (b) the project website[80], and (c) into the project's Zenodo profile, after its publication. Nevertheless, the employed archiving policy is fully aligned with restrictions concerning embargo periods that may be defined by the publishers of these publications, making the latter publicly available in certain repositories only after their embargo period has elapsed.

Dataset name	**MOVING_Data_WP5_2_MOVINGPresentations**
Dataset description	This dataset consists of presentations prepared for reporting MOVING-related scientific work or progress made, in a variety of different events, such as conferences, workshops, meetings, exhibitions, interviews and so on.
Standards and metadata	Most commonly these presentations are in PPT or PDF format. Information related to: (a) the authors, (b) the presenter, (c) the venue and (d) the date of the presentation are also stored in plain text. This dataset has been extended whenever new MOVING presentations were prepared and publicly released. A simple log file of the performed updates of the dataset is maintained by CERTH in the project wiki (hosted by a CERTH server).
Data sharing	The project presentations are made publicly available after their presentation at the venue/event they were prepared for.
Archiving and preservation	The project presentations are publicly available for view and download via the SlideShare channel of the project[81] and the project's Zenodo profile, while links to the presentations on SlideShare were also added to the relevant webpage of the project website[82]. The latter is hosted in a CERTH server that is protected by applying the commonly used security measures for preventing unauthorised access and ensuring that security software is up-to-date with the latest released security

[80] http://moving-project.eu/index.php/publications

[81] http://www.slideshare.net/MOVING_EU

[82] http://moving-project.eu/index.php/presentations

	patches.

Dataset name	**MOVING_Data_WP5_3_MOVINGSoftwareDemosAndTutorials**
Dataset description	This dataset collects information regarding the developed and utilised MOVING technologies. Public video demonstrations, tutorials with instructions of use, documentations as well as links to publicly-released online instances of these technologies are also included.
Standards and metadata	A variety of different formats have been used for storing the necessary information. In particular, video demonstrations are (but not limited to) MP4, AVI or WEBM files, software tutorials and documentations are written in PDF format, online documentations of tools and services is presented in plain text, and presentations are stored in PPT or PDF format. This dataset has been extended whenever new content related to the MOVING developed technologies (e.g. video/web demos, tutorials, documentation) were prepared and publicly released. A simple log file of the performed updates of the dataset is maintained by CERTH in the project wiki (hosted by a CERTH server).
Data sharing	Information related to the developed MOVING technologies, including video demonstrations, documentations, presentations and tutorials with instructions of use, are publicly available supporting the dissemination of the project's activities and the exploitation of the project's outcomes. However, confidentiality control has been applied on each piece of information in order to avoid the release of inappropriate information that could have a negative impact to the project's progress and developments.
Archiving and preservation	Data related to the developed MOVING technologies, tools and applications are archived and made publicly available through the relevant webpage of the project website[83], which is hosted by a CERTH server that is protected by applying the commonly used security measures for preventing unauthorised access and ensuring that security software is up-to-date with the latest released security patches. Moreover, the created video demos and tutorials are also available for view via the

[83] http://moving-project.eu/index.php/tools-and-services

	YouTube channel of the MOVING project[84] and the Videolectures.NET[85] portal.

Dataset name	**MOVING_Data_WP5_4_MOVINGNewsletter**
Dataset description	This dataset comprises the released newsletters for disseminating the activities and the progress made in the MOVING project.
Standards and metadata	The newsletters have been prepared and stored in PDF format, while information regarding their release date provided. This dataset has been extended whenever new project newsletters were publicly released. A simple log file of the performed updates of the dataset is maintained by CERTH in the project wiki (hosted by a CERTH server).
Data sharing	The newsletters of the project were publicly available online right after their official release.
Archiving and preservation	An online archive with open access to the released newsletters of the project is maintained at the relevant webpage of the project website[86], which is hosted by a CERTH server that is protected by applying the commonly used security measures for preventing unauthorised access and ensuring that security software is up-to-date with the latest released security patches.

Dataset name	**MOVING_Data_WP5_5_MOVINGPlatformProspects**
Dataset description	This dataset comprises the opinions expressed by the MOVING project staff and stakeholders concerning the future prospects of technologies, business models and socio-economic trends relevant for the sustainability for the platform. The data gathered within a Delphi survey to be performed in Task 5.2.
Standards and metadata	This dataset is semi-structured: the quantitative opinions are stored in a MySQL database. Each record contains the metadata: an identifier of the respondent, and

[84] https://www.youtube.com/channel/UCLpMLXQQaHDv0CJMG5Sc7mg

[85] http://videolectures.net/moving_videos/

[86] http://moving-project.eu/index.php/newsletters

	date the opinion(s) was expressed. The opinions themselves are the main body of the database consisting of quantitative expressions concerning the state of technology, social, business or economic environment of the MOVING platform at several prescribed moments in the future, the relations between different elements of these environments, and narrative descriptions and justifications. This dataset also contains the results of verifications and trust analysis of expressions contained therein.
Data sharing	Access to the dataset was initially restricted to MOVING partners; it was made publicly available only after a peer-reviewed publication of the analysis of this data by MOVING project staff was published or accepted for publication. The above sharing policy is selected because raw data may be misused by inexperienced readers, specifically by using inappropriate statistical analysis methods, too high or too low confidence values, or inappropriate statistical tests. Therefore, false conclusions may be drawn from the survey data by authors who do not bear full responsibility for gathering and processing the dataset. The use of the data gathered from experts require specialist, carefully and appropriately selected statistical methods, fully compatible with all the survey process. If published results of an inadequate statistical analysis of this dataset are not accompanied by full methodological details, which is a frequent malpractice in scientific community, a proof that the facts are different from those presented by authors from outside of the project may be very difficult or considerably delayed if there were no possibility to refer to the correct analysis published by MOVING authors before.
Archiving and preservation	The dataset is archived at a PBF data server in a MySQL database (structured replies to the survey) and as text notes (free opinions expressed by the survey participants during collaborative sessions supplementing the survey) during the project and its durability period of 5 years after the project end. After the publication, the dataset may also be uploaded into the project's Zenodo profile (the data is exported in .csv or .xls format), and is also stored on MOVING project website as a supplementary material to project publications. A comprehensive description was contained in the publications with data analysis results.

2.5 WP6 Datasets

Dataset name	**MOVING_Data_WP6_1_MOVINGDeliverables**
Dataset	This dataset is composed of the project deliverables that have been prepared and submitted to the EC during the project's lifetime, according to the contractual

description	obligations of the MOVING consortium.
Standards and metadata	These documents are stored in PDF format. For each deliverable we provide: (a) the list of authors, (b) a brief description of its content (i.e. its executive summary), (c) the related WP of the project, and (d) the contractual date for their submission to the EC. This dataset has been extended whenever new deliverables were submitted to the EC. A simple log file of the performed updates of the dataset is maintained by CERTH in the project wiki (hosted by a CERTH server).
Data sharing	The public project deliverables are made publicly available after their submission to the EC, via the project's website, while an abstract was published for the confidential deliverables.
Archiving and preservation	This dataset is maintained on the project wiki and the relevant webpage of the project website[87], both hosted by a CERTH server which is protected by applying the commonly used security measures for preventing unauthorised access and ensuring that security software is up-to-date with the latest released security patches. This webpage grant open access to the PDF file of each listed public deliverable.

[87] http://moving-project.eu/index.php/deliverables

3 Conclusions

The 2nd updated Data Management Plan by the members of the consortium of the MOVING project was presented in this deliverable. This plan involves every dataset that has been collected, processed or generated during the lifetime of the project. In the present document we discussed 53 different datasets, both pre-existing ones (30) and newly-created (23) within MOVING. Most of these datasets are made publicly available, to the benefit of the broader scientific community. The present document represents the 2nd updated dataset-related status and planning of the MOVING project. Though not being a formal deliverable of the project, this update to the Data Management Plan is available via the project's website, similarly to the two previous documents.

References

Beek, W, Rietveld, L, Bazoobandi, R, Wielemaker, J, Schlobach, S (2014). "LOD Laundromat: A Uniform Way of Publishing Other People's Dirty Data", in Proc. of the 13th International Semantic Web Conference, ISWC, Part I, pp. 213-228.

Hoffart, J, Suchanek, F, M, Berberich, C, Weikum, G (2013). "YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia", Artificial Intelligence, Vol. 194, pp.28-61.

Käfer, T, Abdelrahman, A, Umbrich, J, O'Byrne, P, Hogan, A (2013). "Observing Linked Data Dynamics" in Proc. of the 10th Extended Semantic Web Conference, ESWC, Montpellier, France, pp. 213-227.

Konrath, M, Gottron, T, Staab, S, Scherp, A (2012). "SchemEX - efficient construction of a datacatalogue by stream-based indexing of Linked Data", J. Web Sem., 16, 52–58. doi.org/10.1016/j.websem.2012.06.002.

Yang, L, Huang, W, Tan, Ch, L, (2006). "Semi-automatic Ground Truth Generation for Chart Image Recognition", in Proc of the 7th International Workshop (Document Analysis Systems VII), DAS, Nelson New Zealand, pp.324-335.

Appendix

The following informed consent was used for the datasets: MOVING_Data_WP1_3 _Interviews, MOVING_Data_WP1_4_LabUserStudies and MOVING_Data_WP1_5_FocusGroupInterviewTranscript.

Participant Information Sheet – Interview study (TU Dresden, Media Centre)

Approval date *(to be filled out by researchers)*:

This document gives you all information about the conducted study. Please read this sheet carefully and ask questions about anything that you don't understand or want to know in more detail.

1. **Title of the project**

MOVING: TrainInG towards a society of data-saVvy inforMation prOfessionals to enable open leadership INnovation (http://moving-project.eu/)

2. **What is this study about?** *(Project and Study Purpose)*

We are investigating how individuals make sense of data in their daily tasks. By doing so, we will be able to provide training and support, which is tailored to the behaviour, tasks, and information needs of users.

Within this project, we want to conduct the following study:

Aim of the interview study is to evaluate the usage of unstructured data and information as well as strategies for searching and handling data and information in academic work. Therefore, interviews are held with the following two different previously defined academic groups: MA-students and PhD-students in the field of Social Sciences and Humanities. The interviews will be qualitatively analysed. The study takes place as part of a requirement analysis within the forenamed project for the development of an online platform and learning offers to support dealing with unstructured data and information in academia.

3. **Who is running the study** *(Investigators)*

The study is carried out by the following researchers:

Sabrina Herbst, Medienzentrum TU Dresden, Abt. Medienstrategien, sabrina.herbst@tu-dresden.de

Franziska Günther, Medienzentrum TU Dresden, Abt. Medienstrategien, franziska.quenther1@tu-dresden.de

Susann Franke, Medienzentrum TU Dresden, Abt. Medienstrategien, susann.franke@tu-dresden.de

4. **Why have I been invited to participate in the study?** *(Eligibility)*

You have been invited to take part in this research study, because you belong to one of the aforementioned target groups, MA- or PhD-Student in the field of Social Sciences and Humanities.

5. **What would I be asked to do if I take part?** *(Overall Description of Participation)*

During the interview you will be asked to describe your usage behaviour of unstructured data and information, your approach to search and deal with this data/information and the usage of online-based tools in this context.

6. **What is the duration of the research?** *(Length of Participation)*

The interview will take approximately 1-1,5h.

© MOVING Consortium, 2016 1/3

7. **What are the risks associated to the study?** *(Risks of Participation)*

Aside from providing your time, we do not expect any risks/costs associated with taking part in this study.

8. **What are the benefits associated to the study?** *(Benefits of Participation)*

We cannot guarantee or promise that you will receive any direct benefits from being in the study. Your input will be used as for developing a tool/app that provides support for learning data science.

9. **Is there any compensation/payment/incentives?** *(Compensation/Payment/Incentives)*

There will be no compensation.

10. **What happens if I do not want to take part or change my mind?** *(Volunteer Statement)*

It is up to you to decide whether or not to take part. If you decide to participate in the study, you may withdraw from it at any time without giving a reason and with detriment to yourself.

11. **Will the collected information about me be kept confidential?** *(Confidentiality Statement)*

By providing your consent, you agree that we are collecting personal information about you for the purposes of this research study. This information will only be used for the purposes outlined in this Participant Information Sheet. Your information will be stored securely and your identity/information will be kept strictly **confidential**. Study findings may be published, but all data for analysis will be **anonymised**. In reporting on the research findings, we will not reveal the names of any participants. At all times there will be no possibility of you as individuals being linked with the data. Although every effort will be made to protect your identity, there is a risk that your participation (but no individual data) might be identifiable in publications due to the nature of the study and/or the results.

The interview will be electronically recorded and in the context of the study transliterated and scientifically analysed. Analysis and interpretation of the interviews, as well as the citation of interview-sections in scientific publications and presentations will take place anonymised.

12. **What if something goes wrong?** *(Formal complaint about the conduct)*

If you want to make a formal complaint about the conduct of the study, please contact:

Sabrina Herbst, Medienzentrum TU Dresden, Standort Strehlener Straße, 01069 Dresden, sabrina.herbst@tu-dresden.de

Franziska Günther, Medienzentrum TU Dresden, Abt. Medienstrategien, franziska.guenther1@tu-dresden.de

This information sheet is for you to keep

Thank you for reading this information sheet and for considering taking part in this research.

MOVING Project

CONSENT FORM

If you are happy to participate please complete and sign the consent form below

	Tick the box below
1. I confirm that I have read the attached information sheet on the above project and have had the opportunity to consider the information and ask questions and had these answered satisfactorily.	
2. I understand that my participation in the study is voluntary and that I am free to withdraw at any time without giving a reason.	
3. I agree to the use of anonymous quotes.	
4. I agree that any data collected may be passed to other researchers and published in open data repositories.	
5. I am at least 18 years of age.	

I agree to take part in the above project. I understand that I will receive a copy of this form after it has been signed by me and the principal investigator of this research study.

Name of participant	Date	Signature
Name of person taking consent	Date	Signature