So many projects so little time … for DMPs … What do researchers need?

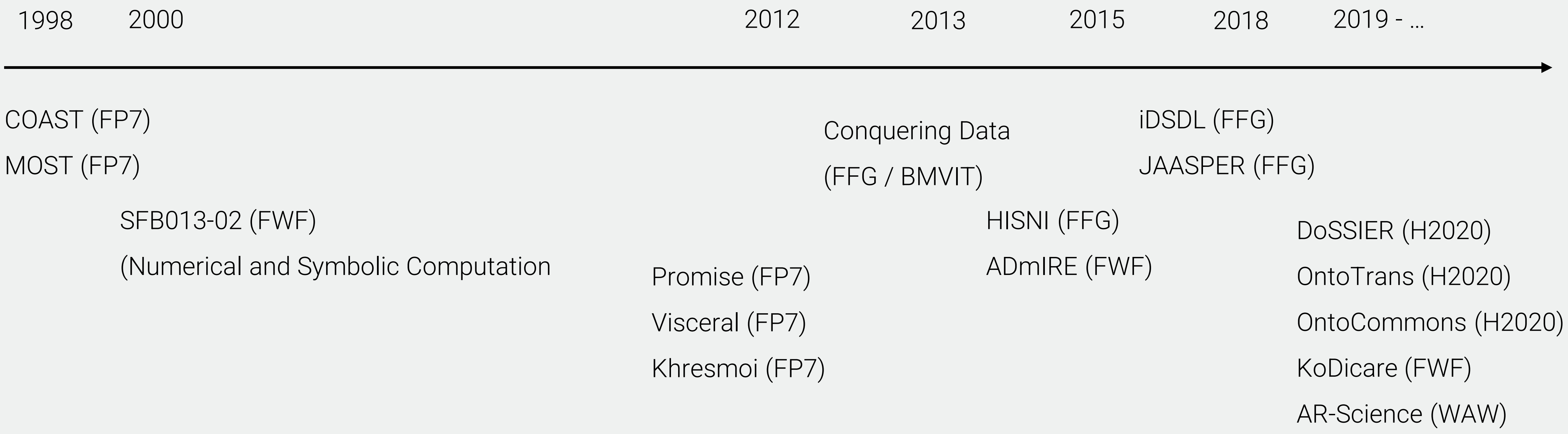
Florina Piroi

Who am I

- Computer Scientist (since 1996)
- MSc In Parallel and Distributed Computing (1998)
- PhD in Symbolic Computation (2004) (Tools for Mathematical Knowledge Management)
- Researcher (since 2000)
    - Data Science
    - Domain Specific Information Retrieval
    - Machine Learning
    - Natural Language Processing
    - …

- Joined the Research Data Management Team at TU Wien couple of weeks ago.

Projects – a timeline



1998

COAST (FP7)

MOST (FP7)

2000

SFB013-02 (FWF)

(Numerical and Symbolic Computation

2012

Promise (FP7)

Visceral (FP7)

Khresmoi (FP7)

2013

Conquering Data

(FFG / BMVIT)

2015

HISNI (FFG)

ADmIRE (FWF)

2018

iDSDL (FFG)

JAASPER (FFG)

2019 - …

DoSSIER (H2020)

OntoTrans (H2020)

OntoCommons (H2020)

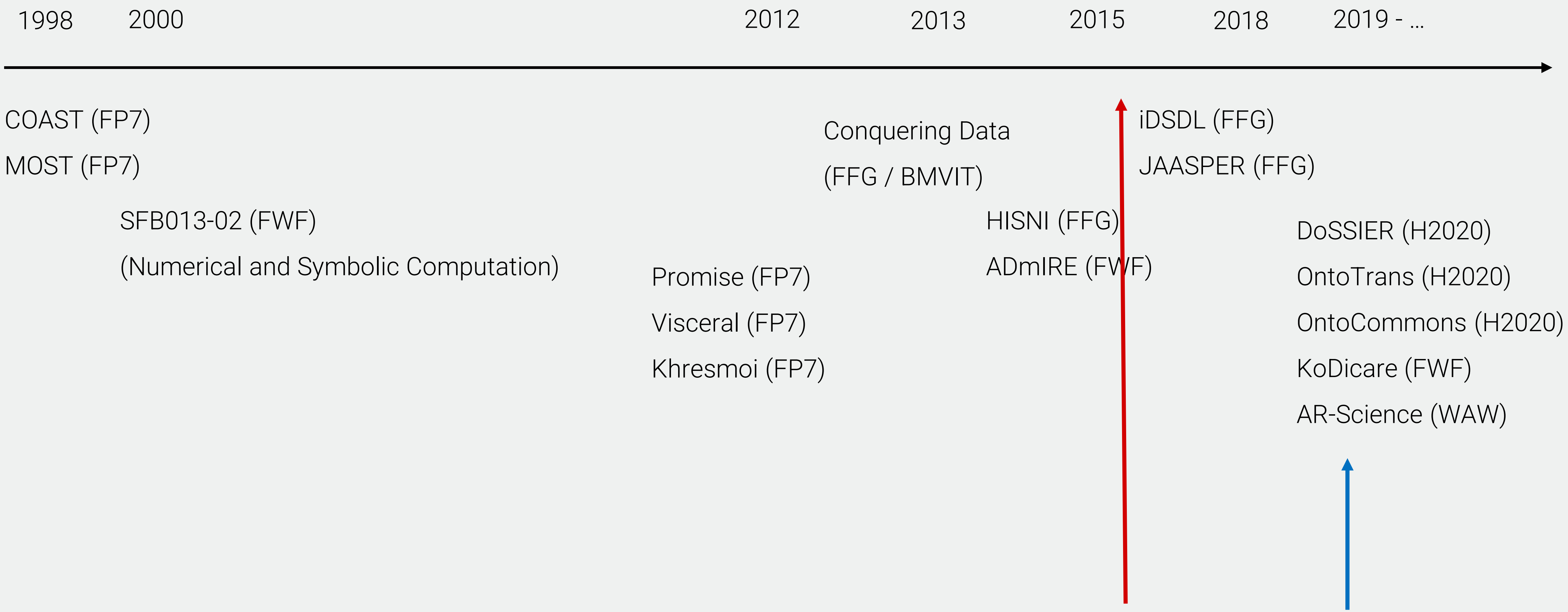
KoDicare (FWF)

AR-Science (WAW)

Data sets

- Medical content
- IP (Patent) documents
- Scientific Articles
- (social network) Logs
- Mathematical content (formulae, proofs, MathML formatted)
- Images (flow-charts, diagrams, chemical formulae, technical drawings
- Industry data (material science, interdisciplinary projects)
- Ontologies & Knowledge Graphs
- …

Projects – a timeline



1998

COAST (FP7)

MOST (FP7)

2000

SFB013-02 (FWF)

(Numerical and Symbolic Computation)

2012

Promise (FP7)

Visceral (FP7)

Khresmoi (FP7)

2013

Conquering Data

(FFG / BMVIT)

HISNI (FFG)

ADmIRE (FWF)

2015

iDSDL (FFG)

JAASPER (FFG)

2018

2019 - …

DoSSIER (H2020)

OntoTrans (H2020)

OntoCommons (H2020)

KoDicare (FWF)

AR-Science (WAW)

DMPs – Selected examples

iDSDL (FFG)

OntoTrans (H2020)

DoSSIER (H2020)

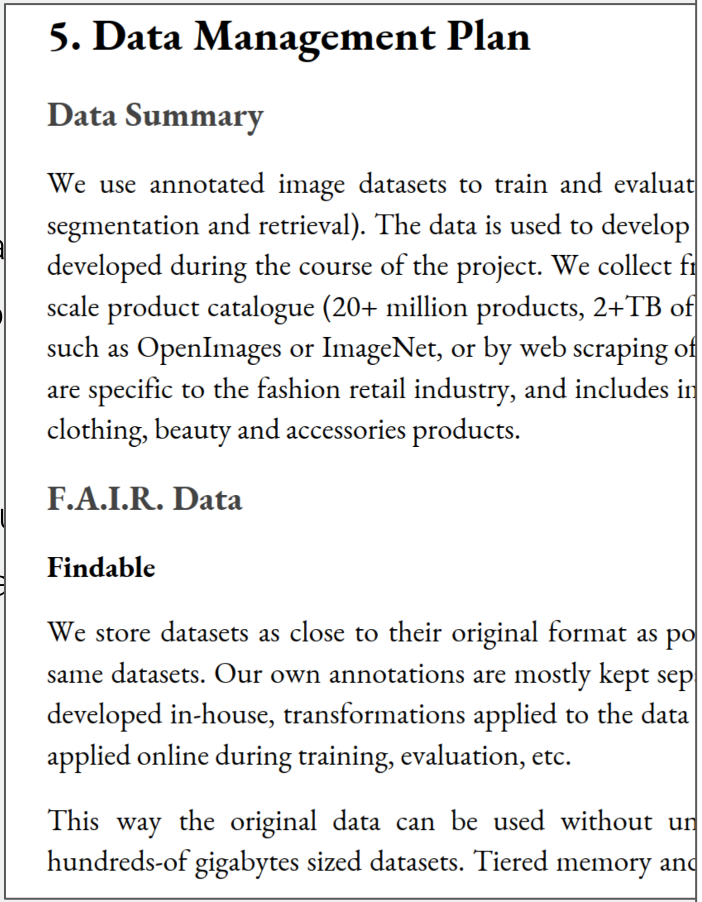
iDSDL (FFG)

- Innovationslehrgang Data Science and Deep Learning
- Transfer of SoTA knowhow on AI, DL, DS to industries in Austria
- 20+ Industrial partners
- 5 teaching modules
- 1 Transfer project per industrial partner
  - Had to also create a DMP

- Expected 20+ DMPs – got lots lots fewer.

iDSDL (FFG)

- Innovationslehrgang Data
- Transfer of SoTA knowho
- 20+ Industrial partners
- 5 teaching modules
- 1 Transfer project per indu
  - Had to also create a



5. Data Management Plan

Data Summary

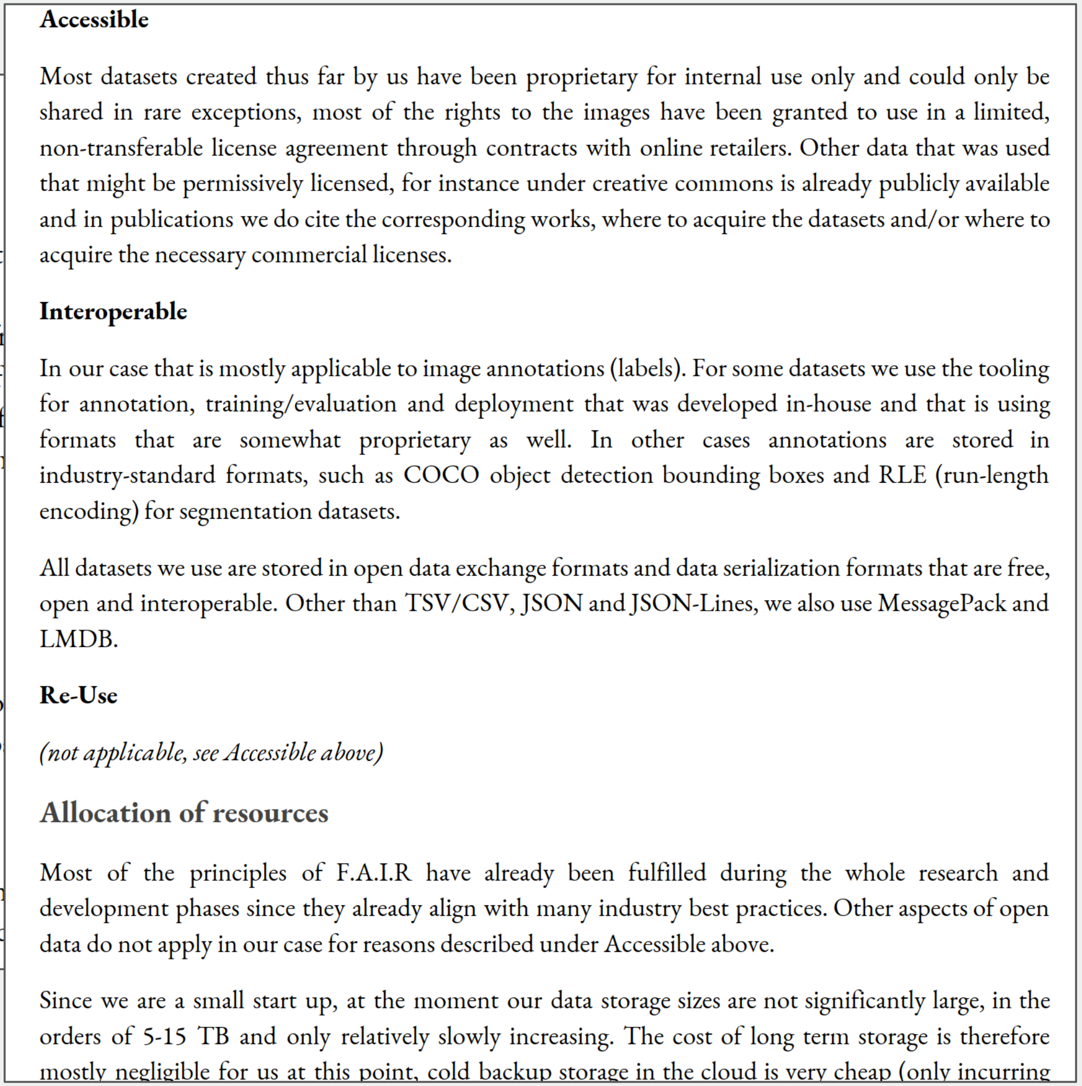
We use annotated image datasets to train and evaluat
segmentation and retrieval). The data is used to develop
developed during the course of the project. We collect fr
scale product catalogue (20+ million products, 2+TB of
such as OpenImages or ImageNet, or by web scraping of
are specific to the fashion retail industry, and includes in
clothing, beauty and accessories products.

F.A.I.R. Data

Findable

We store datasets as close to their original format as po
same datasets. Our own annotations are mostly kept sep
developed in-house, transformations applied to the data
applied online during training, evaluation, etc.

This way the original data can be used without un
hundreds-of gigabytes sized datasets. Tiered memory anc



Accessible

Most datasets created thus far by us have been proprietary for internal use only and could only be shared in rare exceptions, most of the rights to the images have been granted to use in a limited, non-transferable license agreement through contracts with online retailers. Other data that was used that might be permissively licensed, for instance under creative commons is already publicly available and in publications we do cite the corresponding works, where to acquire the datasets and/or where to acquire the necessary commercial licenses.

Interoperable

In our case that is mostly applicable to image annotations (labels). For some datasets we use the tooling for annotation, training/evaluation and deployment that was developed in-house and that is using formats that are somewhat proprietary as well. In other cases annotations are stored in industry-standard formats, such as COCO object detection bounding boxes and RLE (run-length encoding) for segmentation datasets.

All datasets we use are stored in open data exchange formats and data serialization formats that are free, open and interoperable. Other than TSV/CSV, JSON and JSON-Lines, we also use MessagePack and LMDB.

Re-Use

(not applicable, see Accessible above)

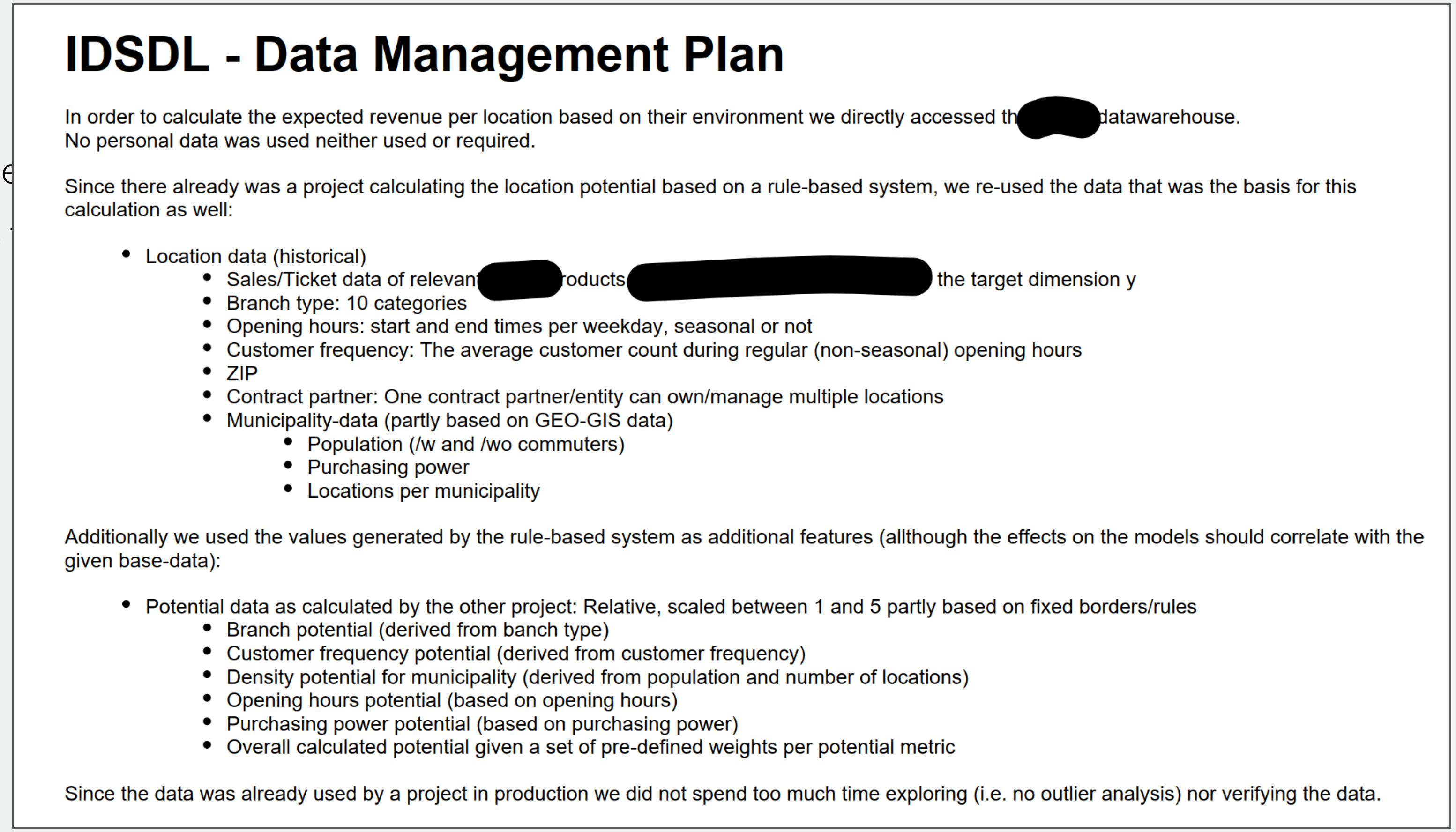
Allocation of resources

Most of the principles of F.A.I.R have already been fulfilled during the whole research and development phases since they already align with many industry best practices. Other aspects of open data do not apply in our case for reasons described under Accessible above.

Since we are a small start up, at the moment our data storage sizes are not significantly large, in the orders of 5-15 TB and only relatively slowly increasing. The cost of long term storage is therefore mostly negligible for us at this point, cold backup storage in the cloud is very cheap (only incurring

iDSDL (FFG)

- Innovationslehrgang Data Science and De
- Transfer of SoTA knowhow on AI, DL, DS
- 20+ Industrial partners
- 5 teaching modules
- 1 Transfer project per industrial partner
  - Had to also create a DMP



IDSDL - Data Management Plan

In order to calculate the expected revenue per location based on their environment we directly accessed th[redacted]datawarehouse.
No personal data was used neither used or required.

Since there already was a project calculating the location potential based on a rule-based system, we re-used the data that was the basis for this calculation as well:

- Location data (historical)
  - Sales/Ticket data of relevan[redacted]roducts[redacted] the target dimension y
  - Branch type: 10 categories
  - Opening hours: start and end times per weekday, seasonal or not
  - Customer frequency: The average customer count during regular (non-seasonal) opening hours
  - ZIP
  - Contract partner: One contract partner/entity can own/manage multiple locations
  - Municipality-data (partly based on GEO-GIS data)
    - Population (/w and /wo commuters)
    - Purchasing power
    - Locations per municipality

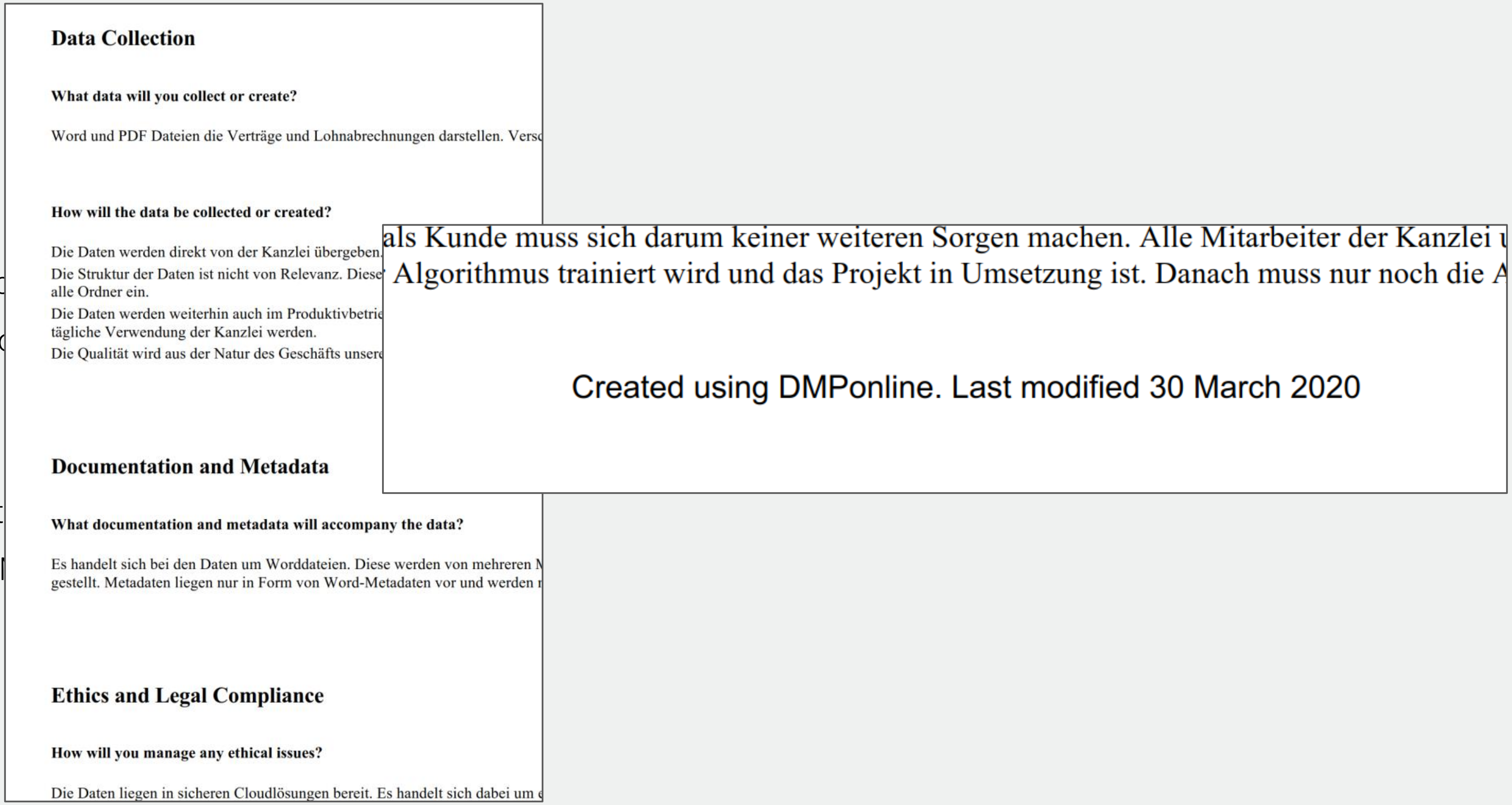
Additionally we used the values generated by the rule-based system as additional features (allthough the effects on the models should correlate with the given base-data):

- Potential data as calculated by the other project: Relative, scaled between 1 and 5 partly based on fixed borders/rules
  - Branch potential (derived from banch type)
  - Customer frequency potential (derived from customer frequency)
  - Density potential for municipality (derived from population and number of locations)
  - Opening hours potential (based on opening hours)
  - Purchasing power potential (based on purchasing power)
  - Overall calculated potential given a set of pre-defined weights per potential metric

Since the data was already used by a project in production we did not spend too much time exploring (i.e. no outlier analysis) nor verifying the data.

iDSDL (FFG)

- Innovationslehrgang Data Sc
- Transfer of SoTA knowhow d
- 20+ Industrial partners
- 5 teaching modules
- 1 Transfer project per indust
  - Had to also create a DI



Data Collection

What data will you collect or create?

Word und PDF Dateien die Verträge und Lohnabrechnungen darstellen. Vers

How will the data be collected or created?

Die Daten werden direkt von der Kanzlei übergeben.

Die Struktur der Daten ist nicht von Relevanz. Diese
alle Ordner ein.

Die Daten werden weiterhin auch im Produktivbetrie
tägliche Verwendung der Kanzlei werden.

Die Qualität wird aus der Natur des Geschäfts unser

Documentation and Metadata

What documentation and metadata will accompany the data?

Es handelt sich bei den Daten um Worddateien. Diese werden von mehreren M
gestellt. Metadaten liegen nur in Form von Word-Metadaten vor und werden n

Ethics and Legal Compliance

How will you manage any ethical issues?

Die Daten liegen in sicheren Cloudlösungen bereit. Es handelt sich dabei um e

als Kunde muss sich darum keiner weiteren Sorgen machen. Alle Mitarbeiter der Kanzlei u
Algorithmus trainiert wird und das Projekt in Umsetzung ist. Danach muss nur noch die A

Created using DMPonline. Last modified 30 March 2020

OntoTrans (H2020)

- Ontology Driven Open Translation Environment
- Material Modelling / Material Sciences

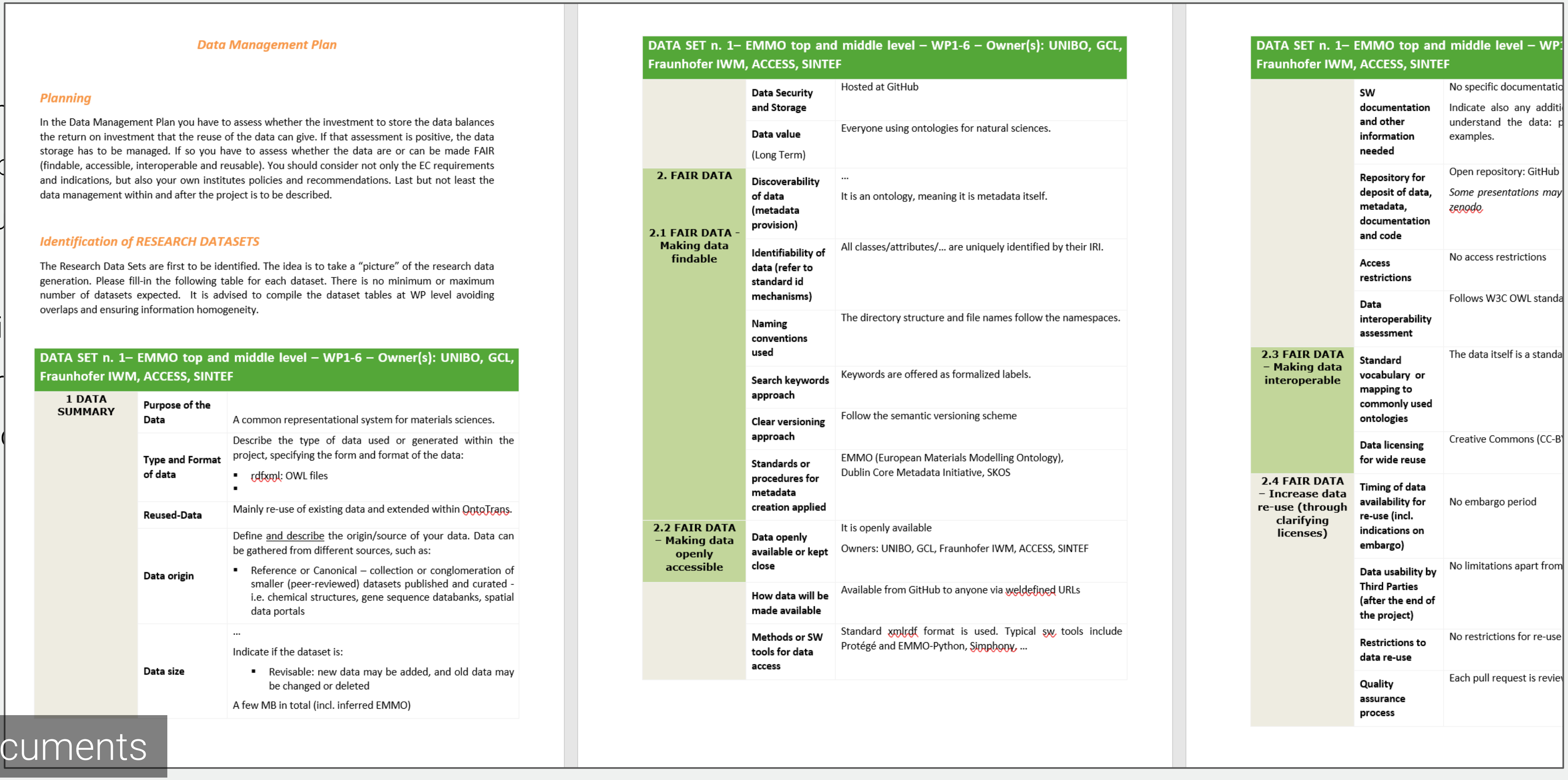
- Ontology creation for specific domains (steel, chemical, prepregs / curing) such that the translation scheme is efficient, cross-domain, adaptable, etc.
- Ontotrans.eu

OntoTrans (H2020)

- DMP created more systematic, now
- EU H2020 template
- Extensive discussions with industrial partners what exactly this plan means in terms of:
    - Access rights and security
    - Disclosure of industrial proprietary data
    - Concept clarification (especially FAIR)
    - …

OntoTrans (H2020)

- DMP created m
- EU H2020 temp
- Extensive discu
  terms of:
  - Access ri
  - Disclosur
  - Concept (



Data Management Plan

Planning

In the Data Management Plan you have to assess whether the investment to store the data balances the return on investment that the reuse of the data can give. If that assessment is positive, the data storage has to be managed. If so you have to assess whether the data are or can be made FAIR (findable, accessible, interoperable and reusable). You should consider not only the EC requirements and indications, but also your own institutes policies and recommendations. Last but not least the data management within and after the project is to be described.

Identification of RESEARCH DATASETS

The Research Data Sets are first to be identified. The idea is to take a "picture" of the research data generation. Please fill-in the following table for each dataset. There is no minimum or maximum number of datasets expected. It is advised to compile the dataset tables at WP level avoiding overlaps and ensuring information homogeneity.

DATA SET n. 1– EMMO top and middle level – WP1-6 – Owner(s): UNIBO, GCL, Fraunhofer IWM, ACCESS, SINTEF		
1 DATA SUMMARY	**Purpose of the Data**	A common representational system for materials sciences.
	Type and Format of data	Describe the type of data used or generated within the project, specifying the form and format of the data: ▪ rdfxml: OWL files ▪
	Reused-Data	Mainly re-use of existing data and extended within OntoTrans.
	Data origin	Define and describe the origin/source of your data. Data can be gathered from different sources, such as: ▪ Reference or Canonical – collection or conglomeration of smaller (peer-reviewed) datasets published and curated - i.e. chemical structures, gene sequence databanks, spatial data portals
	Data size	… Indicate if the dataset is: ▪ Revisable: new data may be added, and old data may be changed or deleted A few MB in total (incl. inferred EMMO)

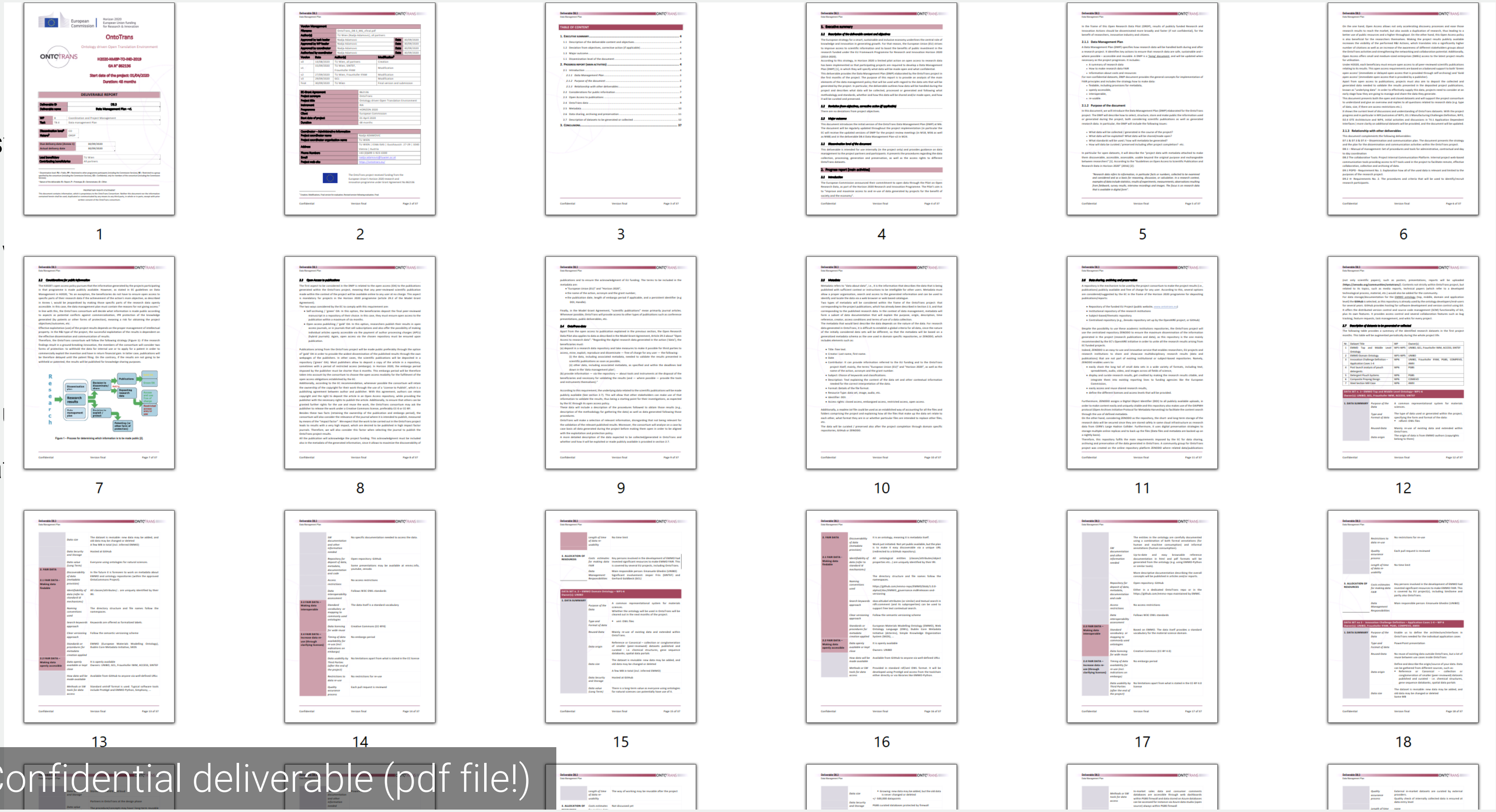
DATA SET n. 1– EMMO top and middle level – WP1-6 – Owner(s): UNIBO, GCL, Fraunhofer IWM, ACCESS, SINTEF		
	Data Security and Storage	Hosted at GitHub
	Data value (Long Term)	Everyone using ontologies for natural sciences.
2. FAIR DATA **2.1 FAIR DATA - Making data findable**	**Discoverability of data (metadata provision)**	… It is an ontology, meaning it is metadata itself.
	Identifiability of data (refer to standard id mechanisms)	All classes/attributes/… are uniquely identified by their IRI.
	Naming conventions used	The directory structure and file names follow the namespaces.
	Search keywords approach	Keywords are offered as formalized labels.
	Clear versioning approach	Follow the semantic versioning scheme
	Standards or procedures for metadata creation applied	EMMO (European Materials Modelling Ontology), Dublin Core Metadata Initiative, SKOS
2.2 FAIR DATA – Making data openly accessible	**Data openly available or kept close**	It is openly available Owners: UNIBO, GCL, Fraunhofer IWM, ACCESS, SINTEF
	How data will be made available	Available from GitHub to anyone via weldefined URLs
	Methods or SW tools for data access	Standard xmlrdf format is used. Typical sw tools include Protégé and EMMO-Python, Simphony, …

DATA SET n. 1– EMMO top and middle level – WP1 Fraunhofer IWM, ACCESS, SINTEF		
	SW documentation and other information needed	No specific documentatio Indicate also any additi understand the data: p examples.
	Repository for deposit of data, metadata, documentation and code	Open repository: GitHub *Some presentations may zenodo*
	Access restrictions	No access restrictions
	Data interoperability assessment	Follows W3C OWL standa
2.3 FAIR DATA – Making data interoperable	**Standard vocabulary or mapping to commonly used ontologies**	The data itself is a standa
	Data licensing for wide reuse	Creative Commons (CC-B
2.4 FAIR DATA – Increase data re-use (through clarifying licenses)	**Timing of data availability for re-use (incl. indications on embargo)**	No embargo period
	Data usability by Third Parties (after the end of the project)	No limitations apart from
	Restrictions to data re-use	No restrictions for re-use
	Quality assurance process	Each pull request is revie

About 8 such documents

OntoTrans (H2020)

- DMP created more sys
- EU H2020 template
- Extensive discussions
  terms of:
  - Access rights
  - Disclosure of indi
  - Concept clarifica



All collected into one Confidential deliverable (pdf file!)

DoSSIER (H2020)

- Domain Specific Systems for Information Retrieval
- Dossier-project.eu
- MSCA ITN/ETN
    - 15 subprojects (15 PhDs)
    - Lots! of data!

- Instructions sent to students with tables to fill
- Top-down approach

DoSSIER (H2020)

- Domain Specific Systems for Information Retrieval
- Dossier-project.eu
- MSCA ITN/ETN
  - 15 subprojects (15 PhDs)
  - Lots! of data!

- Instructions sent to students with tables to fill
- Top-down approach

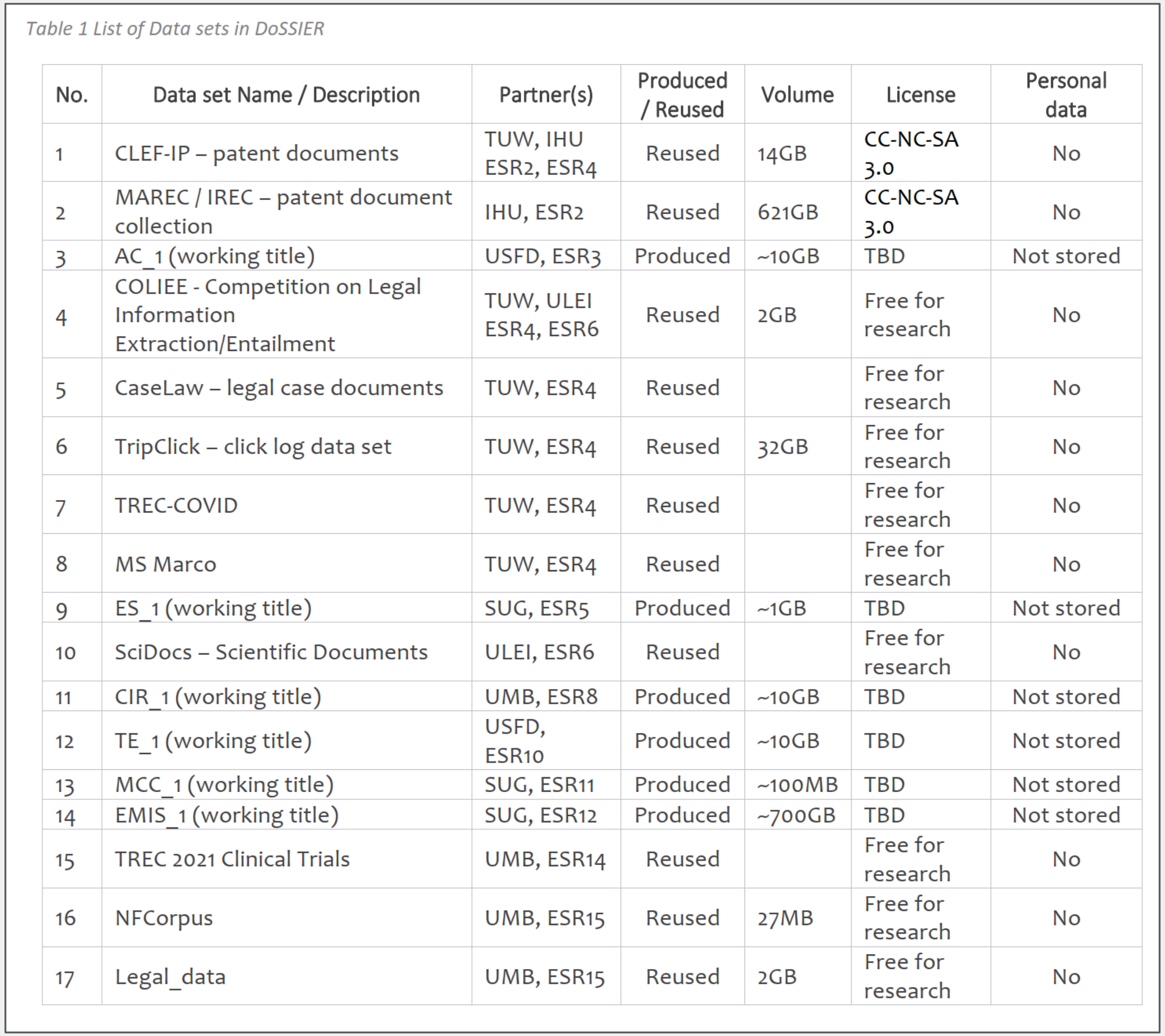


Table 1 List of Data sets in DoSSIER

No.	Data set Name / Description	Partner(s)	Produced / Reused	Volume	License	Personal data
1	CLEF-IP – patent documents	TUW, IHU ESR2, ESR4	Reused	14GB	CC-NC-SA 3.0	No
2	MAREC / IREC – patent document collection	IHU, ESR2	Reused	621GB	CC-NC-SA 3.0	No
3	AC_1 (working title)	USFD, ESR3	Produced	~10GB	TBD	Not stored
4	COLIEE - Competition on Legal Information Extraction/Entailment	TUW, ULEI ESR4, ESR6	Reused	2GB	Free for research	No
5	CaseLaw – legal case documents	TUW, ESR4	Reused		Free for research	No
6	TripClick – click log data set	TUW, ESR4	Reused	32GB	Free for research	No
7	TREC-COVID	TUW, ESR4	Reused		Free for research	No
8	MS Marco	TUW, ESR4	Reused		Free for research	No
9	ES_1 (working title)	SUG, ESR5	Produced	~1GB	TBD	Not stored
10	SciDocs – Scientific Documents	ULEI, ESR6	Reused		Free for research	No
11	CIR_1 (working title)	UMB, ESR8	Produced	~10GB	TBD	Not stored
12	TE_1 (working title)	USFD, ESR10	Produced	~10GB	TBD	Not stored
13	MCC_1 (working title)	SUG, ESR11	Produced	~100MB	TBD	Not stored
14	EMIS_1 (working title)	SUG, ESR12	Produced	~700GB	TBD	Not stored
15	TREC 2021 Clinical Trials	UMB, ESR14	Reused		Free for research	No
16	NFCorpus	UMB, ESR15	Reused	27MB	Free for research	No
17	Legal_data	UMB, ESR15	Reused	2GB	Free for research	No

DoSSIER (H2020)

- Domain Specific Systems for Information Retriev
- Dossier-project.eu
- MSCA ITN/ETN
  - 15 subprojects (15 PhDs)
  - Lots! of data!

- Instructions sent to students with tables to fill
- Top-down approach

- Each data set described separately, in an additional table

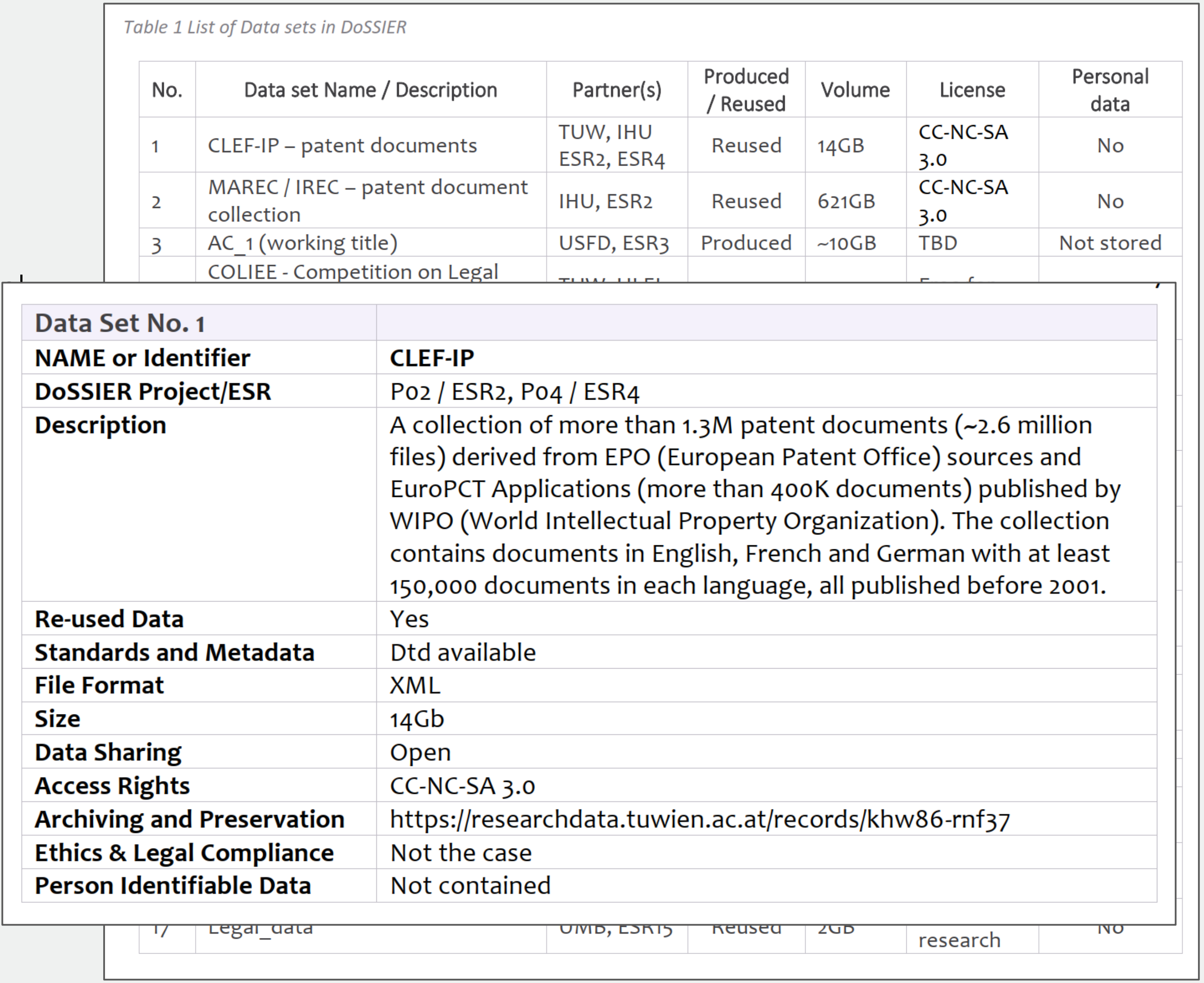


Table 1 List of Data sets in DoSSIER

No.	Data set Name / Description	Partner(s)	Produced / Reused	Volume	License	Personal data
1	CLEF-IP – patent documents	TUW, IHU ESR2, ESR4	Reused	14GB	CC-NC-SA 3.0	No
2	MAREC / IREC – patent document collection	IHU, ESR2	Reused	621GB	CC-NC-SA 3.0	No
3	AC_1 (working title)	USFD, ESR3	Produced	~10GB	TBD	Not stored
	COLIEE - Competition on Legal					
17	Legal_data	UMB, ESR15	Reused	2GB	research	No

Data Set No. 1	
NAME or Identifier	**CLEF-IP**
DoSSIER Project/ESR	P02 / ESR2, P04 / ESR4
Description	A collection of more than 1.3M patent documents (~2.6 million files) derived from EPO (European Patent Office) sources and EuroPCT Applications (more than 400K documents) published by WIPO (World Intellectual Property Organization). The collection contains documents in English, French and German with at least 150,000 documents in each language, all published before 2001.
Re-used Data	Yes
Standards and Metadata	Dtd available
File Format	XML
Size	14Gb
Data Sharing	Open
Access Rights	CC-NC-SA 3.0
Archiving and Preservation	https://researchdata.tuwien.ac.at/records/khw86-rnf37
Ethics & Legal Compliance	Not the case
Person Identifiable Data	Not contained

Lessons Learned

Did you notice?

- All just pdf files – static
- No follow up on them (that I am aware of)
- Not retrievable for statistics
- Very different information (template dependent)
- “why” not clear enough in the community
- KISS – researchers don’t want overhead related to data management (but no way out – how do we tell them that?)
- And: often researchers (in my domain) release some data, somewhere (e.g. Kaggle, hugginface), little overlap with DMP (tool-ed or pdf-ed)

What we’d like

- Simple guidelines (actually available!)
- Small overhead
- Early introduction to the whole ecosystem of data management
- Proof of benefit
- Institutional support (advisors, community, etc)