Topic Modelling workflow report

Big Data approaches for improved monitoring of research and innovation performance and assessment of the societal impact in the Health, Demographic Change and Wellbeing Societal Challenge

Project acronym: Data4Impact

Grant Agreement no: 770531

Deliverable: D4.1

<table>
<thead>
<tr>
<th>Deliverable information</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Deliverable number and name</td>
<td>D4.1 Topic Modelling workflow report</td>
</tr>
<tr>
<td>Due date</td>
<td>Month 12</td>
</tr>
<tr>
<td>Delivery</td>
<td>Month 12</td>
</tr>
<tr>
<td>Work Package</td>
<td>WP4</td>
</tr>
<tr>
<td>Lead Partner for deliverable</td>
<td>ARC</td>
</tr>
<tr>
<td>Author</td>
<td>Omiros Metaxas (ARC)</td>
</tr>
<tr>
<td>Reviewers</td>
<td>All partners</td>
</tr>
<tr>
<td>Dissemination level</td>
<td>Confidential</td>
</tr>
<tr>
<td>Version</td>
<td>0.2</td>
</tr>
</tbody>
</table>
Table 1. Document revision history

<table>
<thead>
<tr>
<th>Issue Date</th>
<th>Version</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>03/08/2018</td>
<td>0.1</td>
<td>First draft</td>
</tr>
<tr>
<td>26/08/2018</td>
<td>0.2</td>
<td>Final draft</td>
</tr>
</tbody>
</table>
Disclaimer

This document contains description of the Data4Impact project findings, work and products. Certain parts of it might be under partner Intellectual Property Right (IPR) rules so, prior to using its content please contact the consortium coordinator for approval.

In case you believe that this document harms in any way IPR held by you as a person or as a representative of an entity, please do notify us immediately.

The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content.

This publication has been produced with the assistance of the European Union. The content of this publication is the sole responsibility of Data4Impact consortium and can in no way be taken to reflect the views of the European Union.

The European Union is established in accordance with the Treaty on European Union (Maastricht). There are currently 28 Member States of the Union. It is based on the European Communities and the member states cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice and the Court of Auditors.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 770531.
Contents

Disclaimer................................................................................................................................................. 4

Contents ......................................................................................................................................................... 5

Abbreviations .................................................................................................................................................. 6

Summary ........................................................................................................................................................ 7

1. Introduction and Explanatory Remarks.................................................................................................... 8

2. Description of methodology....................................................................................................................... 8
   2.1 Data collection and pre-processing ...................................................................................................... 10
   2.2 Information Enrichment & Network Incorporation ........................................................................... 11
   2.3 Topic Modelling ................................................................................................................................... 12
   2.4 Topic based information retrieval ....................................................................................................... 12
   2.5 Trends & Similarities Calculation ....................................................................................................... 13
   2.6 Exploration, Visualization and Analysis ............................................................................................. 13
      2.6.1 Research map (similarity graph)................................................................................................. 13
      2.6.2 Chord diagrams ........................................................................................................................... 15
      2.6.3 Trend Diagrams ............................................................................................................................ 15

3. Related data workflow, techniques, tools and metrics ............................................................................. 16
   3.1 Techniques and algorithms ................................................................................................................. 16
   3.2 Quantitative analysis & topic related metrics: .................................................................................. 17
   3.3 Description of the related data flow, inputs/outputs and tools ............................................................ 18

4. Indicative work plan & risk analysis ......................................................................................................... 20
   4.1 Work plan ........................................................................................................................................... 20
   4.2 Risk analysis ...................................................................................................................................... 21

5. References .................................................................................................................................................. 22
Abbreviations

**PTM** – Probabilistic Topic Modelling

**EC** – European Commission

**ARC** – Athena Research & Innovation Center

**UoA** – University of Athens

**WP** – Work Package

**WT** - Wellcome Trust

**NIH** - National Institute of Health
Summary

This document provides the description of the Topic Modelling Workflow including all related tools, services, visualization, data inputs and data outputs (results). Such workflow, utilizes and extends already existing services initially developed as part of OpenAIRE’s platform to mine throughput data (e.g., publications and patents) related to the Health, Demographic Change and Wellbeing Societal Challenge. Therefore, this document describes a) the innovative platform for information (entity) extraction and automated multi-dimensional analysis of all scholarly or patent related content (including textual information, meta-data and extracted entities / links) based on probabilistic topic modelling and other machine (deep) learning techniques and b) the related analysis methodology as well as new indicators that will provide useful insight and timely intelligence of systematic research in E.U.
1. Introduction and Explanatory Remarks

Deliverable 4.1 is part of the Work Package 4 “Topic modelling of throughput data”. It describes the tools, metrics and a related well-defined methodology to explore, model and analyse research and patent documents promoting innovation, policy making and impact assessment. In more detail, WP4 objectives can be summarized as follow:

- identify active areas of research: discover hidden themes (topics)
- understand what is actually produced (quantitative outcome analysis): calc topic distributions per document / entity (e.g., project, author, call)
- analyze active research areas on several dimensions (e.g., compare geographic regions, funders, etc.) assessing dynamic topic activation per entity over the years, topic exclusivity etc
- discover clusters and communities, assess research collaboration: topic-based similarity analysis
- identify emerging research areas: topic-based trend analysis
- assess coverage, identify gaps or new challenges: (e.g., assessing dynamic topic activity over the years per funder -that quantifies produced output- for each specific challenge, comparing EU funded research vs global research map, analysing topic exclusivity on several dimensions)
- map topics to specific challenges described in each call (semi-automated process)
- evaluate the role of funding (assessing topic activity for each entity involved, e.g., authors, before and after the call)
- assess the impact of research in the society using new indicators (research topic trend index, weight, exclusivity index, trend vs exclusivity etc)

WP4 is split into following tasks:

- Task 4.1 Training and development of models for the Health, Demographic Change and Wellbeing Societal Challenge [lead Athena RC; UoB and PPMI contributing]
- Task 4.2 Topic modelling of publication data related to the Health, Demographic Change and Wellbeing Societal Challenge: static and trend/dynamic analyses [lead Athena RC; PPMI contributing]
- Task 4.3 Topic modelling of patent data related to the Health, Demographic Change and Wellbeing Societal Challenge: static and trend/dynamic analyses [lead Athena RC; PPMI contributing]

As shown in WP4 data analysis flow diagram in fig.1, Task 4.1 focusses on the design and the implementation of the topic modelling analysis workflow including data collection, data filtering, enrichment and pre-processing. Tasks 4.2 & 4.3 focus on the purpose specific model building (4.2 on publications & 4.3 on patents) and the analysis of the related results. Hence, we plan to build specific topic models providing comprehensive reports, visualizations, indexes and metrics that will provide useful insight and timely intelligence of systematic research worldwide. Proposed analysis will incorporate multi-dimensional, spatio-temporal modelling capturing trends (emerging areas) and comparing systematic research and/or modelling associations among different entities (e.g., geographic regions, funding programmes etc) as well as topic-based similarity analysis graphs among different entities (e.g., grants).

2. Description of methodology

We utilize and extend the already existing dynamic Multi-View Topic Modelling platform, developed by ATHENA RC, which has already been incorporated within OpenAIRE’s inference workflows to provide useful insight and timely intelligence of-funded-systematic research in the E.U. and/or worldwide. Such platform consists of several tools, metrics, services and visualizations targeting information (entity) extraction, semantic annotation and, most importantly, automated multi-dimensional analysis based on an innovative multi-view probabilistic topic modelling engine. The latter, analyses scientific publications, patents and other related information as well as varied additional side or extracted information (e.g,
authors, venues, grants, semantic annotations, bio-entities) and links (e.g., citation network) aiming to alleviate the impact of information overload in Research & Innovation (R&I). Based on this platform, and following a well-defined methodology that covers all related steps starting from data collection, data extraction, network analysis and semantic information enrichment to multi-view topic modelling, expert curation and validation, topic-based similarity detection and trend analysis, we are able to explore, model and analyse research empowering innovation, policy making and impact assessment.

As mentioned above we utilize and extend the already existing dynamic Multi-View Topic Modelling platform, developed by ATHENA RC, which has already been incorporated within OpenAIRE’s inference workflow to explore, model, analyse and visualize systematic research in E.U related to the Health, Demographic Change and Wellbeing Societal Challenge. Such platform consists of several tools, metrics, services and visualizations targeting information (entity) extraction, semantic annotation and, most importantly, automated multi-dimensional analysis based on an innovative multi-view probabilistic topic modelling engine. The latter analyses all associated textual content (e.g., publications, deliverables, reports, work programmes, related policy documents and patents) and related side information (e.g., metadata, semantic annotations, links) in order to identify underlying thematic information (i.e., low-dimensional multi-view latent representations named “topics”) and corresponding overlapping clusters/communities combining all disparate information sources aiming to alleviate the impact of information overload in Research & Innovation (R&I). The generated topics serve as the means to connect the dots between technical advances, concepts, people, organizations, funding, or even between different kind of entities (e.g., bio-terms) within and across scientific disciplines and technical domains.

In addition, they are subsequently used in information mapping and retrieval mechanisms for external data collection processes (e.g., to collect/filter social or company data, expand and refine search and disambiguate or filter related results and entities), as well as in trend, network/community and similarity analysis related tasks.

To achieve a more coherent and interpretable result, we use a multi-view probabilistic topic modelling engine developed by ATHENA RC which jointly analyses textual and related side information and identifies hidden themes (topics) that characterize them. Multiple entity types (‘views’) can help to explain each other and the discovered multi-view topics are better correlating with human judgment of topical quality. In addition, we achieve higher coverage, uncovering concepts not resolved by traditional, textual only topic models.

The diagram in fig.1 describes the proposed iterative analysis workflow of the whole WP4 that captures two distinct interrelated processes: (i) topic identification (T4.1), and (ii) topic modelling based information retrieval and analysis (T4.2 & T4.3). Each of these processes is related to several sub-tasks, each one of them addressing different aspects of the analysis. As the diagram shows, it is important to emphasise that in several of these sub-tasks, human (expert) intervention will be needed for the curation, validation and evaluation of the results, as well as in the categorisation and topic labelling process. Finally, all multi-view topics are automatically semantically annotated using well-defined ontologies (DBPedia for general purpose labelling, and MeSH for biomedical annotation).
In T4.2 and T4.3 we plan to build and analyse more than one models (e.g., one for publications and project reports focusing on funded research (e.g., from European Commission, National Institute of Health, Welcome Trust, National Funders etc), one on patent data and a final one that will try to map, analyse and compare health related research worldwide -at global level- as described by all collected data including the whole PubMed Corpus -incorporating all Open Access publications and publications where we have the at least the abstract-. For each of the above models, we will have multiple -incremental-iterations. In the final iteration we may also include project-related content from web sites, company data, or content related to EU directives and regulations.

It is important to notice that all related flow has been already utilized in several EU funded tenders and projects. Therefore, both the topic modelling engine and several information extraction / annotation services and web-based visualizations are in place supporting the whole flow. Nevertheless, further developments and extensions are being implemented in order to release an integrated platform able to seamlessly support the whole flow.

### 2.1 Data collection and pre-processing

Data4Impact collects and analyse the following data (already described in D1.4 Data Management Plan):

- input data (programmes, projects; national and EU level);
- throughput data (publications + their full texts; patents + full texts);
- output & impact data (company data; monitoring data on finalised EU projects; online/social media blog/health-related forum data; clinical guideline and policy documents/data).
Although, our analysis in WP4 focuses on throughput data (publication and patent data) related to the Health, Demographic Change and Wellbeing Societal Challenge, our goal is to incorporate and compare findings from related multidimensional analysis on company, project report and social data that is performed in WP5. The following table summarise the data that are collected and analysed in WP4.

**Table 1. Overview of data that are collected and analysed in WP4**

<table>
<thead>
<tr>
<th>Data Description &amp; Type</th>
<th>Method</th>
<th>Data origin</th>
<th>Data collection purposes (brief explanation)</th>
<th>Data utility, Outputs of the Analysis</th>
<th>Lead partner(s) handling data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Publications:</strong> text i.e., abstract and full text, existing or extracted metadata i.e., funding info, semantics (MeSH, Wikipedia), citations, authors, venues, publication date.</td>
<td>Using APIs (publications), Web crawling (CORDIS), data dump (EPO)</td>
<td>Publications: OpenAIRE, PubMed, Patents: European Patent Office (EPO)</td>
<td>Topic Modelling based analysis that will provide useful insight and timely intelligence of systematic research in E.U</td>
<td><strong>Active areas of research:</strong> discover hidden themes (topics)</td>
<td>ATHENA</td>
</tr>
<tr>
<td><strong>Patents:</strong> Full text and metadata</td>
<td></td>
<td></td>
<td></td>
<td><strong>Outcome quantitative analysis:</strong> calc topic distributions per document / entity (e.g., project, author, call)</td>
<td></td>
</tr>
<tr>
<td><strong>Project Data:</strong> Project description, executive summary, metadata</td>
<td>Project Data: OpenAIRE &amp; CORDIS, National Funding Agencies</td>
<td></td>
<td></td>
<td><strong>Active research areas on several dimensions / comparisons (e.g., geographic regions, funders, etc.):</strong> Topic based exclusivity analysis</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Clusters and communities:</strong> topic based similarity analysis / collaboration assessment</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Emerging research areas:</strong> topic based trend analysis</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Coverage, gaps or new challenges:</strong> compare funded research, identify exclusive but emerging or under represented topics</td>
<td></td>
</tr>
</tbody>
</table>

In more detail, we collect and analyse data from various sources and in various formats including:

**Publications and related metadata:** We collect publicly available information from OpenAIRE, and PubMed (based on standard APIs) including all associated textual content (i.e., abstracts and full text whenever available) and related side information (e.g., metadata like venue, publication date etc, semantic annotations like MeSH, funding information, citation and authorship & affiliation links).

**Project outputs about publicly funded projects:** We collect publicly available information from CORDIS (using web crawling techniques) including project reports, abstracts, evaluation summary, research programme and other related metadata. We also incorporate information related to projects for non E.U. finders including Wellcome Trust (WT), National Institute of Health (NIH) and SWEDISH National funded projects using OpenAIRE or National APIs.

**Patents (IPR):** Textual content (abstract and full text), authors and related metadata from European Patent Office (under specific license).

After data collection, we apply standard pre-processing techniques for text mining (e.g., tokenization, stemming, stop word removal).

### 2.2 Information Enrichment & Network Incorporation

Besides already existing metadata (venue, keywords, MeSH terms etc) we enrich related textual content with additional side information (‘views’) such as funding information, and semantic terms. For semantic
annotation based on DBPedia terms. Our DBPedia semantic annotator is using the open source DBPedia Spotlight annotator and the DBPedia Sparql Endpoint. In addition, we may utilise several other NER tools for bio-entity annotation. Subsequently, an innovative and efficient way to incorporate structural network information within topic models is used, based on a network wordfication (propositionalization) technique that efficiently combines network ranking with topic modelling.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Text</td>
<td>topic, latent, lda, document, news, dirichlet, probabilistic, mining, mixture, allocation, articles, generative, semantic, temporal, word, topical, corpus, pisa, bayesian</td>
<td>mapreduce, big, scalability, hadoop, scalable, analytics, cluster, datasets, cloud, map, intensive, jobs, massive, queries, google, job, machines, node, computations, mining, hdfs, hive, workloads</td>
</tr>
<tr>
<td>Citations (titles) (ranked list of citation net nodes)</td>
<td>“Dynamic topic models”, “Topics over time”</td>
<td>“A comparison of approaches to large-scale data analysis”, “Pig latin”, “Mesos”, “DryadLINQ”, “PREGEL”, “CIEL”, “Improving MapReduce performance in heterogeneous environments”, “MapReduce Online”, “MapReduce Merge”</td>
</tr>
<tr>
<td>Keywords</td>
<td>topic modeling, latent dirichlet allocation, latent semantic analysis, generative model, text mining,</td>
<td>Map-Reduce, big data, hadoop, cloud computing, distributed computing, data analytics, machine learning, parallel processing</td>
</tr>
</tbody>
</table>
| Venues       | SIGKDD, WSDM, CIKM                                                                   | SIGMOD, BigSystem, CloudCP, EUROSEC, EUROSYS, ...

| DBpedia (Wikipedia) terms | Topic_model, Statistical_model, Multinomial_distribution, Probabilistic_latent_semantic_analysis, Mixture_model, Latent_Dirichlet_allocation, Bayesian_inference | MapReduce, Apache_Hadoop, Big_data, Data_processing, Scalability, Data_analysis, Data_management, NoSQL, Data_science, Data_parallelism, Partition_of_a_set, Bigtable, Apache_HBase, Dryad |
| HDP_LDA (text only baseline) | topic, sentiment, opinion, reviews, comments summarization, mining, latent, forums, sentences, news, summary, lda, topical, textual | similarity, index, indexing, mapreduce, indexes, hadoop, inverted, indices, queries, databases, indexed, similarities, searching, measure, map, collections, scalable, measures |

Figure 2: Examples of two multi-view topics demonstrating interpretability and coherence. Proposed multi-view topic model (above) analyzes 6 ‘views’ that include text, relational (citation net), side (metadata) and extracted (DBpedia annotations) information (left column). Inferred topics are described using a ranked list of terms for each view (including citations & DBpedia links). Text only HDP-LDA baseline (bottom) cannot uncover specific topics like “Topic Modeling” or “Cloud computing & Big Data Analytics” mixing either text mining, summarization, opinion mining and topic modelling related words in the first, or similarity search, indexing and MapReduce related words in the second.

2.3 Topic Modelling

Utilizing our multi-view probabilistic topic modeling engine, we jointly analyze textual and related side information and identify hidden themes (topics) that characterize them. Multiple entity types (‘views’) can help to explain each other and the discovered multi-view topics that are more coherent and interpretable, hence, better correlating with human judgment of topical quality. In addition, we achieve higher coverage, uncovering concepts not resolved by traditional, textual only topic models. We also assign topics to publications (calculating related topic proportions). Topic related metrics (described in the next section) are used for topic qualitative and quantitative analysis as well as model selection. Experts can help in topics validation, assessment, categorization and labeling process.

2.4 Topic based information retrieval

Identified latent thematic information (i.e., ‘topics’) as well as related semantic terms are used to expand and refine search and information retrieval from additional resources (WEB, blogs, EU portals, project/company web sites, etc). Most informative terms, tokens and topics are identified based on a new metric that assesses the discriminative strength of a token or of a topic. Such metric can be further be
used to evaluate the coverage of the underlying concepts across one specific dimension (e.g. Projects). Results are validated by our analysts.

2.5 Trends & Similarities Calculation

Based on inferred topics per publication we are able to calculate topic proportions for other related entities such as topics per grant, per author, per research area, per funder, etc. Then we calculate similarities among those entities based on their topic distributions. In addition, we calculate topic trends per different entities (topics distributions over time). The topic-based similarity (e.g., among projects) allows to explore, characterize and summarize connections and analogies between projects. Through a quantification of the similarity of pairs of projects, it is possible to cluster related projects and make special and temporal comparisons.

2.6 Exploration, Visualization and Analysis

All data including collected data, metadata, extracted entities and links, classifications, topics, topic proportions and related topic based similarities, distributions and trends are stored in one relational database and can be accessed and explored through multiple querying interfaces and APIs.

Content-based visualization of the scientific results is a challenging task mainly due to the richness and diversity of such content, especially when seen from the aspect of the research administrator of the funder, whose main interest is of assessing research impact or of strategy/policy making. Effective visualizations can provide illustrations of the content distribution for particular facets of interest: funding schemes, countries, universities, authors, scientific fields, etc. Therefore, we create several interactive data driven WEB based visualizations (e.g. similarity research graph/map, trend analysis diagrams) that provide useful insights and can help experts and policy makers to identify emerging, exclusive or common topics, assess research timeliness, discover hidden patterns, similarities, structure & communities discover hidden patterns, similarities, structure & communities.

Below we show some examples:

2.6.1 Research map (similarity graph)

Based on the above mentioned PTMs, we are implementing an interactive WEB based graph that focus on capturing and demonstrating topic-based similarities between projects/grants, publications or topics, and identifying groupings and relations among them. Our interactive GUI borrows and extends several ideas from state of the art related topic modeling visualization projects such as (Gretarson, et al., 2012) (Bruce, Talley, Burns, Newman, & LaRowe, 2009).

Our goal is to produce a Research / Funding interactive map creating a node-link graph where nodes are projects/grants, publications or authors and edges capture topic based similarities between them. In addition, node radius corresponds to the number of publications and colour to related research area (e.g. PEOPLE, ICT, HEALTH, etc.). In order to place nodes, we are based on a Force Directed Layout algorithm (Jakobsen & Lee, 2012) with simple constraints (Dweyer, 2009) and collision detection.
Figure 3: Similarity graph of FP7 projects (1)

Figure 4: Similarity graph of FP7 projects (2)
2.6.2 Chord diagrams

Chord diagrams analyze connections among more generic entities (e.g., FP7 (sub) research areas). This way a policy maker can see cross-disciplinary projects and research areas, interconnections among them, etc.

Figure 5: Relations analysis of FP7 research areas chord diagram

2.6.3 Trend Diagrams

Such diagrams assess or compare topic activation over time. This way one can identify emerging topics, or old fashion topics. Trend diagrams can either be related to the whole corpus, or be entity specific, i.e., topic activation per research area, funder, author or venue.

Figure 6: Topic activation over time (1)
3. Related data workflow, techniques, tools and metrics

3.1 Techniques and algorithms

The whole analysis is based on an intelligent and scalable probabilistic framework for mixed-membership multi-view learning on Text Augmented Heterogeneous Information Networks (TA-HINets) that are composed of interconnected entities (e.g., publications) characterized by free text attributes, relational information (e.g., links in citation networks), and other side information (e.g., labels, taxonomies, authors or other meta-data). In more detail, we have developed an innovative non-parametric Multi-View topic model (MV_HDP\(^1\)), which extends the well-established generative probabilistic topic models (PTMs) based on hierarchical Bayesian analysis and Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, Latent Dirichlet allocation, 2003) (Blei D., 2012) (Steyvers & Griffiths, 2006) (Teh, Jordan, M., Beal, & Blei, 2006), that have been successfully used to mine textual content revealing

---

\(^{1}\) O.Metaxas, Y.Ioannidis, Multi-View Topic Modelling on Text-Augmented Heterogeneous Information Networks (under review for TKDD journal)
the underlying structure of a document collection. Incorporating a novel Interacting Pólya Urn scheme (IUM) to model per-document topic distribution, hence combining interaction and reinforcement, MV_HDP addresses the following major challenges:

a) multiple views sharing statistical strength,

b) estimation and adaptation to the extent of correlation between them

d) scalable inference on massive, real world datasets handling models with thousands of topics.

Also, we have incorporated a novel approach to efficiently incorporate relational information within topic models based on a network wordification pre-processing step and a new metric that assesses the discriminative strength of a topic to further evaluate the quality of the inferred topics and the coverage of the underlying concepts.

The intuition behind Latent Dirichlet Allocation (LDA) is that a single document exhibits more than one ‘topics’ in different proportions. A topic is defined as a probability distribution over a fixed vocabulary (terms). A document is modeled as a probability distribution over topics. LDA is a generative probabilistic model that defines a joint probability distribution over both observed and hidden variables. Here the observed variables are the words of the documents and the hidden variables are the document/topic proportions and topic/words distributions.

Recently many models have been proposed that extend classical Latent Dirichlet Allocation (LDA) or its non-parametric sibling Hierarchical Dirichlet Process (HDP) in one of the following directions: a) relational topic models, trying to tie links and text (Chang & Blei, 2009), (Wang, Silva, Willett, & Carin, 2011) b) multi-view(modal) topic models that mainly target to tie image and text (Du, Lu, Dunson, & Carin, 2009) (Jia, Salzmann, & Darrell, 2011) and c) supervised topic models that combine text data and related meta-data. Methods in the latter direction include Statistical Entity Models (Newman, Chemudugunta, & Smyth, 2006), sLDA (Blei & McAuliffe, 2007), DMR (Mimno & McCallum, 2008), DCNT (Kim & Sudderth, 2011), Author Topic model (AT) (Rosen-Zvi, Griffiths, Steyvers, & Smyth, 2004), Labeled LDA (LLDA) (Ramage, Hall, Nallapati, & Manning, 2009) and Partially Labeled PLDA (Ramage, Manning, & Dumais, Partially Labeled Topic Models for Interpretable Text Mining, 2011) Flat-LDA, PL-DP (Rubin, Chambers, Smyth, & Steyvers, 2011).

Our work provides significant advantages over such supervised methods as it neither assumes the existence of a set of labels on every document nor explicitly assigns every word to a specific label. Our focus is not on predicting labels (classify documents) or assign topics to labels but on inferring well-defined multi-view topics that best describe the underlying concepts. Thus, MV_HDP can be characterized as a multi-view model. Compared to existing multi-view models, MV_HDP is unique in the ability of directly modeling any number of views, learning and adapting to the degree of correlation between them. In addition, instead of directly modeling links, we efficiently incorporate relational information through a network wordification pre-processing step that aligns the disparate representations of text and network data. Finally, MV_HDP is non-parametric and able to scale on real world datasets handling millions of documents on a single PC.

3.2 Quantitative analysis & topic related metrics:

Model checking and evaluation for topic modeling is an open research question despite significant related work (Wallach, Murray, Salakhutdinov, & Mimno, 2009; Mimno, Wallach, Talley, Leenders, & McCallum, 2011; Chang, Boyd-Graber, Gerrish, Wang, & Blei, 2009) that exists. In every case, such metrics (especially those assessing model fitting) do not necessarily correlate with human judgment (Chang, Boyd-Graber, Gerrish, Wang, & Blei, 2009; Mimno, Wallach, Talley, Leenders, & McCallum, 2011), hence, topic inspection and human evaluation are always necessary. We use the following two metrics for automatically evaluating a model’s fitting and topic quality:

- **Model fitting (LL):** To evaluate model fitting and convergence for the training data, we use the log likelihood (LL) of the data given the model logP(w|T) (Griffiths & Steyvers, 2004). Predictive negative LL estimates are normalized by word counts determining perplexity scores (Blei, Ng, & Jordan, 2003). We calculate LL for every different view.
Topic Coherence (Coh): To evaluate topic coherence, we use an existing metric (Mimno, Wallach, Talley, Leenders, & McCallum, 2011) relies upon word co-occurrence statistics. It is very similar to pointwise mutual information (PMI) but focuses on the conditional probability of each word given each of the higher-ranked words in a topic.

Besides the above mentioned model and topic evaluation/selection metrics, we are also calculating additional metrics that can be used to assess the qualitative characteristics of a topic such as:

- **Topic weight** – How often a topic is encountered (i.e. topic strength):
  - common vs rare topics

- **Trend index** – capture topic strength fluctuation over time. We calculate the ratio of the average topic weight for two user defined periods (old and new):
  - emerging vs old-fashion vs all time classic topics

- **Exclusivity index (Discriminative weight)** – evaluate the discriminative strength of a topic in one specific dimension (e.g., publications, grants, funders, journals etc). The more evenly a topic is distributed across this dimension, the less exclusive it is. Such metric reveals many interesting aspects of topics usage
  - High (exclusive): topic is only mentioned within a small number of entities (e.g., grants, journals or conferences), i.e., targeted distribution (with spikes)
  - Low (shared): topic spread evenly over many entities, i.e., flatter distribution

By combining all above mentioned metrics as well as related graphs one can identify potential gaps (e.g., emerging but rare topics only addressed by a small number of grants), assess coverage, or identify research areas where a more focused call is required (e.g., emerging topics evenly spread across many calls).

3.3 Description of the related data flow, inputs/outputs and tools

The following tables and diagrams provide a very detailed description for the whole Topic Modelling workflow, including related data inputs and outputs and tools.

Table 2: High-level data flow

<table>
<thead>
<tr>
<th>Input data description</th>
<th>Tools</th>
<th>Output data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>T4.1 Content enrichment &amp; Topic Modelling Flow: Athena RC</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Project abstracts, Reports Related publications (fulltext) and metadata Patent related data (fulltext)</td>
<td>ATHENA RC SciTopic: Multi-View Probabilistic Topic Modelling engine (Java) Semantic annotators, NER tools</td>
<td>Enriched documents with Semantic Annotations or links to extracted entities Multi-View Topics (Thematic Areas) Topic Per document / Patent Topic qualitative and quantitative analysis</td>
</tr>
<tr>
<td><strong>T4.2 &amp; T4.3 Analysis of the results: ATHENA RC &amp; PPMI</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inputs from previous step</td>
<td>Informative terms, topics (Java app / SQL script) Trends &amp; Similarities calculation (Java app / SQL Scripts) Interactive WEB Based Visualizations (D3.js)</td>
<td>Informative terms General topic trends / trends per different entities (e.g., project, Programme) Similar entities and related entity similarity graph: Similar projects / Research Areas Associations of projects to E.U. Directives / Legislation</td>
</tr>
</tbody>
</table>
### Table 3: Related sources and technical details

<table>
<thead>
<tr>
<th>Input data description</th>
<th>Analysis</th>
<th>Output data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input data will be extracted from:</td>
<td>The whole analysis will be based on a Java based Multi-View Topic modelling engine from ATHENA RC. Additional semantic annotation (DBPedia Spotlight &amp; NERs tools will be used)</td>
<td>Output data will be in <strong>Backend</strong>: SQL Database format (we are currently using PostgreSQL); <strong>Schema</strong>: see below; <strong>Size</strong>: ~100 GB; <strong>Access protocols</strong>: ODBC; <strong>Exchange representation format</strong>: JSON/CSV</td>
</tr>
<tr>
<td>• EC CORDA database or Cordis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• OpenAccess PubMed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• OpenAIRE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Patents Full text</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All data will be imported in a PostgreSQL DB</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

**Figure 9: Topic Modelling workflow and related tools, modules and technologies**

Most of the above-mentioned modules have already been implemented in the context of OpenAIRE. Probabilistic topic modeling is based on a scalable and efficient Multi-View, semantic topic model (MV_HDP2) which is able to share statistical strength among textual attributes and other related side information taking into account the degree of correlation between them, model and incorporate relational information and scale to massive, real world datasets. The whole engine is built upon MALLET3 toolkit (a Java-based package for statistical natural language processing, document classification, clustering, topic modeling, information extraction, and other machine learning applications to text). Almost any relational database (accessible through JDBC) can be used for data collection, harmonisation and output of the results (currently either PostgreSQL or SQLite are used). File based data access and export is supported. Network Analysis / Wordification is built upon the open-source GraphChi4 disk-based large-scale graph computation system (which is the single node Turi's GraphLab Create engine 5). DBPedia semantic annotator is utilizing the open source DBPedia Spotlight annotator 6 as well as DBPedia.

---

2 O.Metaxas, Y.Ioannidis, Multi-View Topic Modelling on Text-Augmented Heterogeneous Information Networks (under submission)
3 [http://mallet.cs.umass.edu/](http://mallet.cs.umass.edu/)
4 [https://github.com/GraphChi](https://github.com/GraphChi)
5 [https://turi.com/](https://turi.com/)
Sparql Endpoint7. Additional, text mining services like jtopia8 keyword/key phrase extractor are used for other iterelated tasks like automatic topic entitlement.

4. Indicative work plan & risk analysis

4.1 Work plan

As mentioned above WP4 is split into following tasks:

- **Task 4.1 Training and development of models for the Health, Demographic Change and Wellbeing Societal Challenge** [lead Athena RC; UoB and PPMI contributing]
- **Task 4.2 Topic modelling of publication data related to the Health, Demographic Change and Wellbeing Societal Challenge: static and trend/dynamic analyses** [lead Athena RC; PPMI contributing]
- **Task 4.3 Topic modelling of patent data related to the Health, Demographic Change and Wellbeing Societal Challenge: static and trend/dynamic analyses** [lead Athena RC; PPMI contributing]

Below we show an indicative work plan for all above analysed subtasks.

<table>
<thead>
<tr>
<th>Sub-task</th>
<th>Detailed activities</th>
<th>Star/ end dates</th>
<th>Output / Deliverable / milestone</th>
<th>Interfaces with</th>
<th>Leading partner</th>
<th>Inputs needed from other partners</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Task 1.1. Training and development of models</strong></td>
<td>Data collection, filtering and pre-processing</td>
<td>Apr 2018 / July 2018</td>
<td>Relevant publications, project reports &amp; patent data (WP2, WP3) Relevant company and social data (WP5)</td>
<td>WP2, WP3, WP5</td>
<td>ATHENA RC</td>
<td>Relevant data (publications, patents, project reports, social and company data etc)</td>
</tr>
<tr>
<td><strong>Information enrichment &amp; Network incorporation</strong></td>
<td>May 2018 / Sept. 2018</td>
<td>Semantic annotations and new entities identification</td>
<td>-</td>
<td>ATHENA RC</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td><strong>Topic modelling workflow</strong></td>
<td>June 2018 / Oct 2018</td>
<td>Specific topic models on publication and patent data D4.1 Proposed Topic Modelling workflow report (M10) Year1 milestone: topic modelling flow up and running</td>
<td>-</td>
<td>ATHENA RC</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td><strong>Task 4.2 Topic modelling of publication data: static and trend/dynamic analysis</strong></td>
<td>Topic exploration, curation, entitling and validation</td>
<td>Septem 2018/ Jan 2018</td>
<td>• Curated, entitled and categorized topics • Most informative topics, keywords, semantic terms Year1 milestone: 1st iteration on gathered publication data</td>
<td>WP5, WP6</td>
<td>ATHENA RC</td>
<td>• Thematic Expertise</td>
</tr>
<tr>
<td><strong>Calculate trends and similarities</strong></td>
<td>Dec 2018/</td>
<td>• Topic based trends (general and per</td>
<td>WP6</td>
<td>ATHENA RC</td>
<td>• Thematic Expertise</td>
<td></td>
</tr>
</tbody>
</table>

---

7 [http://dbpedia.org/sparql](http://dbpedia.org/sparql)
8 [https://github.com/srijiths/jtopia](https://github.com/srijiths/jtopia)
4.2 Risk analysis

The following risks were identified and assessed. We have also developed mitigation strategies. However, should new risks arise, they will equally be assessed and mitigation strategies implemented.

Risk 1: Data collection (project outputs, OpenAIRE data, Patents data): EC does not agree to share the CORDA data, legal agreement with OpenAIRE, access to patent full text data

A bulk of CORDA data can be substituted with open-source CORDIS data (related crawler has already been implemented). We already work on forming a legal business agreement with OpenAIRE. We intend to buy full text patent data.

Risk 2: Difficulties curating and analysing the results.

We intend to use domain experts that will validate, curate and analyse produced results. In addition, we will repeatedly tweak the model specifications and analysis algorithms.
5. References


