IRFD Project Description · DFF-projektbeskrivelse

RApid GeNeration And pReservation of infOrmation and Knowledge from big data (big data RAGNAROK)

Nikola Vasiljević

Problem statement:

Researchers need to be able to use big data resources efficiently, but heretofore have had very few tools to systematically record their analysis processes, which can be machine-actionable. Such 'analysis preservation' is central to verification and reuse of research results, but in many fields, it remains the responsibility of individual researchers, using ad-hoc methods—if it is attempted at all. As a result, researchers employing big data resources are invariably faced with exploration 'from scratch'; for example, re-analyzing entire datasets in search of features, which were actually identified by preceding inquiries. Consequently, research investment is frequently duplicated and opportunities for building new knowledge upon previous experience are lost.

This proposal addresses these problems by applying new data annotation techniques and standards to systematically preserve information and knowledge obtained in the course of both human and machine interactions with big data. Annotations will capture information about features identified in datasets together with machine-actionable relations between Digital Objects (DO) that led to the identification of the annotated feature, such as datasets, queries, processing code, etc. The annotation system will be the core of a comprehensive, domain-agnostic platform for long-term preservation of DOs and knowledge corpora. The proposed solution will be built and demonstrated around a wind energy use case and generalized for other domains.

State-of-the-art:

A recent survey, reported by Forbes [1], showed that collecting, organizing, and cleaning data (i.e., data preparation) accounts for 80% of researcher's time, whereas the remaining 20% of their time is spent on the actual data analysis. In the same survey, 76% of researchers stated that data preparation is the least favorite part of their work. Consequently, the underlying DOs are difficult and inefficient to re-use, both by other researchers and the DO authors themselves after a certain period of time (reusability).

The FAIR principles, introduced in 2016 [2], provided a set of general guiding principles for improving the infrastructure supporting the reuse of research data (by making data Findable, Accessible, Interoperable and Reusable). The implementation of FAIR principles is an important first step towards making the work with data more efficient, but tools are required to exploit the full potential of the principles.

Considering the ever-increasing volume of data, the ability to search and identify data subsets, which contain specific already identified features, is becoming essential. Annotations of data have emerged as a new research paradigm to serve this purpose, creating significant opportunities for improving discovery and consequent re-use of research investments. For example, Image-Net [3], a database of 14 million organized and labeled (i.e., annotated) images, was instrumental for the computer-vision breakthroughs that came since 2009 [4]. This annotated database of images allows discovery and aggregation of specific features of interest (e.g., pictures containing tiger cats [5]).
Annotations are either embedded with data or they are stand-off DOs. The latter type is preferred since it leaves the annotated data intact and it will be used in the project. Currently, under the auspice of Research Data Alliance (RDA), standards to create discoverable stand-off annotation are under development [6]. Nevertheless, the possibilities, which data annotations offer, have been under-utilized by the research community. In the proposed project, we intend to make tools that will unleash the data annotation capabilities, ultimately saving researchers time, while accelerating findability, accessibility, interoperability and reusability of research data.

Method:

To support the project goal, we will create a platform, which will provide following high-level services: (1) browsable repository for DOs, (2) registration of DOs persistent identifiers (PID), (3) data annotation infrastructure and (4) generation of information and knowledge corpora. To build the platform, we will use existing state-of-the-art software technologies based on standards and focus upon integrating the technologies effectively (see Figure 1). This will allow us to concentrate on adapting annotation tools to numerical wind energy datasets and building robust annotation services for the broader community. Our principal components will comprise a workflow infrastructure **freizo**[7] and a repository **Invenio**[8]. The **freizo** workflow, an open source infrastructure developed by Data Futures, has been used internationally for more than a decade. Particularly, it has been used in the **hasdai** project [9] to create the corpora of the **hasdai** network of **Invenio** repositories. **Invenio** is an open source repository management framework developed by CERN that underpins Zenodo [10], the global catch-all repository for scientific research endorsed by the G7 and funded by the European Commission's OpenAIRE program.
We will also employ open source annotation instruments for 2D imagery, which are already available, such as **Mirador** [11] and **Universal Viewer** [12], which will immediately allow annotation of visual data plots, prior to the development of machinic annotation tools (for application directly to numerical data) based on the Machine Learning (ML) algorithms such as **OpenCV** [13] . We will use the GitHub repository to manage development of the platform. Also, we will adopt RDA-based FAIR and PID schemes. Additionally, we will employ the metadata

methodologies of Schema.org and DataCite on the existing wind energy-related controlled vocabularies and taxonomies [14] making them available to the global community.

As stated earlier, the **big data RAGNAROK** project will be based around an **Invenio** repository (see repository in Figure 1), created as a node of the **hasdai** network thereby speeding up the integration of our research environment and providing a route to industrial production readiness for later trials. In a pilot project during April and May 2019, a proof-of-concept **freizo** workflow using a DTU wind lidar dataset has been created [15] and it already supports annotation of data plots and ORCID [16] authentication of contributors. In the first phase of the **big data RAGNAROK** project, we will use the pilot project results to build the entire platform. The first step will be to create a repository information and knowledge corpus from these preliminary annotations and consolidate the corpus using PIDs (connect the corpus to underlying DOs), followed by automation of the corpus deposit to Zenodo and DTU Data [17] and its versioning from the annotation workflow (**freizo**). Besides wind energy datasets, which will reside on our repository, we will also support datasets residing on external IT infrastructure. However, all metadata including PIDs, as well as annotations will reside in our repository. We will use standards-based interfaces to connect datasets and the related annotation workflows to the repository. This will make our approach equally applicable to numerical datasets in other research domains, and also that external annotation tools can be employed for the future enrichment of our repository. Integration with **hasdai** will support discovery by the global community using our taxonomies, as well as providing high security and accessibility for our corpora by developing institutional guarantee policies with organizations such as **sciCORE** [18]. Further, we will support repository interoperability and provide an automated deposit mechanism to Zenodo and DTU Data with DOIs for DOs (datasets, code, annotations, etc.) in our repository during the first phase of the project.

In the second phase of the project, we will focus on the development of tools for machinic annotation (i.e., automatic annotation using ML techniques) of wind energy datasets; integrating the repository with front-end data processing and yielding speed-ups and cost reduction, which will radically change the reusability and financial landscape of wind energy data. Software arising from these activities will be made public and archived via a dedicated GitHub repository. We will make an automated deposit of the GitHub repository to Zenodo and DTU Data. We will also establish a compliance-checking service for wind energy datasets to be uploaded to the repository, following recommendations of RDA's FAIR Data Maturity Model WG. In the third phase of the project, we will develop a browsable access service for the wind energy community using the repository, providing unprecedented support for navigation of our datasets (see Corpora in Figure 1) and also precipitate pilot projects (the phase four of the project) with communities generating large potentially annotatable numerical datasets in other research domains. During the lifetime of the

project, our infrastructure will be tested against at least two high-quality data collections from the wind energy domain: New European Wind Energy Atlas data collection [19] and the wind characteristics database [20].

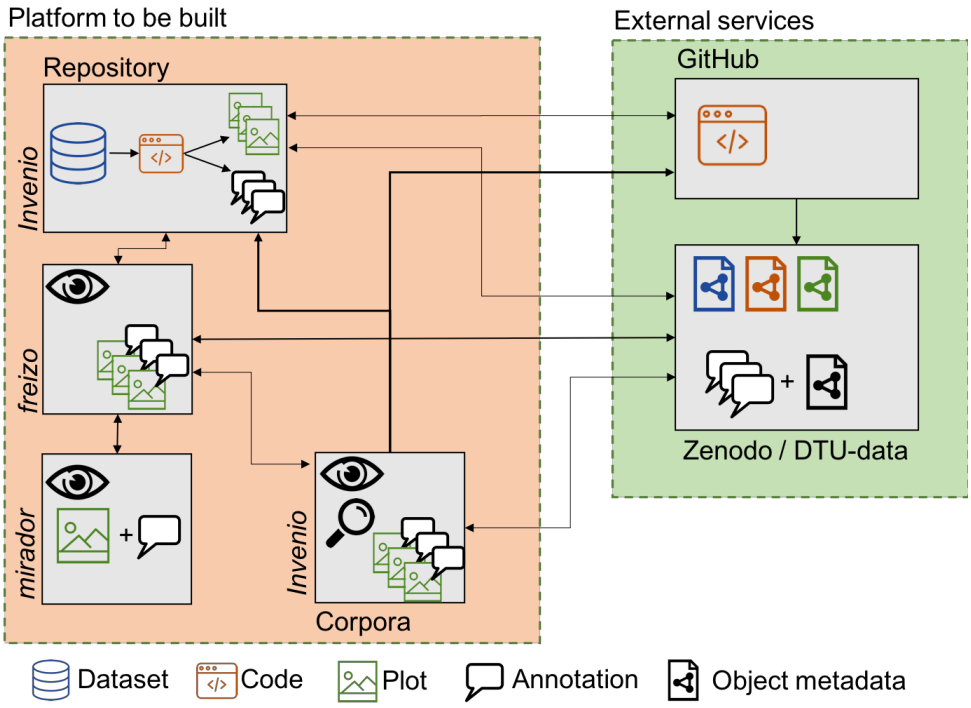

Figure 1. System architecture represented on example of single dataset

Scientific and societal perspective and relevance of project:

Through radical innovation, the project will spearhead a new era of generation and preservation of information and knowledge from research data by developing a set of services (based on existing open source solutions) and making them seamlessly integrated. This will dramatically boost the efficiency in research, since the previous records of interaction with data and all the underlying DOs will be preserved, interconnected and made publically available. Overall, the proposed solutions will significantly improve the impact of research investments by making them simpler for reuse and accessible to everyone, thus enabling open research and citizen science. The continuously updated information and knowledge corpora linked to the underlying DOs will facilitate and secure long-term reuse of research investments. It will make these investments robust, especially when the research staff involved in their creation are no longer present. The corpora will be a 'one-stop-shop' for data required to build and train AI algorithms, empowering open innovations resulting in the development of cutting-edge data-driven services.

Research plan and practical framework of the project:

The project (Figure 2) is expected to start in April 2020 and to last for 3 years. The project is split into four phases. The first phase, with a duration of 18 months, will be focused on platform

development. The second phase, with a duration of 9 months, will be focused on the development of tools for automatic annotation of datasets. The third phase, with a duration of 6 months, will be focused on the browsable access service development. The fourth phases, with the duration of 6 months, will be dedicated to organizing and carrying out pilot projects within Technical University of Denmark (DTU) where the solutions developed in the first three phases of the project will be applied to numerical datasets from departments other than wind energy. The lead applicant will spend several short-term visits at Data Futures in the first three phases of the project. DTU Wind Energy and Data Futures are main responsible for the development of all proposed services, while during the development DTU Library will act as a steering committee making sure that the services are applicable across the entire DTU. DTU Library is main responsible for organizing pilot projects within DTU and compiling guidelines for the platform use and its integration with repositories and FAIR data management. In this activity, DTU Wind Energy and Data Futures will act as a support. We will organize two workshops. An introductory workshop about the platform will take place prior the start of pilot projects. A post-pilot projects workshop will take place following the completion of the pilot projects where the pilot projects results will be presented, which will be an opportunity to discuss follow-up projects and dissemination activities. Besides several physical meetings (e.g., during the workshops and short-term stays) we will have monthly video meetings while daily conversions will take place at a mattermost team site hosted at DTU Wind Energy [21]. The project plan contains following milestones: M1 - Project plan and preliminary results presented at the 17th RDA plenary meeting, M2 - Platform developed and connected to external services (Zenodo, DTU Data, etc.), M3 - Platform presented at 19th RDA plenary meeting, M4 - Machinic annotation services developed and integrated in the repository, M5 - Browsable access service developed and interfaced to the repository, M6 – Guidelines how to use the platform published, M7 – Paper presenting the developed platform submitted to Scientific Data Journal, M8 - Pilot projects completed, M9 - Paper summarizing results of pilot projects submitted to CODATA Journal.

	2020				2021				2022				2023			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
WS1 Platform							M2									
WS2 Machinic annotation										M4						
WS3 Browsable access											M5					
WS4 Pilot projects													M8			
WS5 Management																
WS6 Dissemination					M1				M3		M6	M7		M9		

Figure 2. Project Gant chart: M – milestone and WS – work stream. Each year split to quarters.

References:

[1]https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#2e7ed7c46f63

[2] https://www.nature.com/articles/sdata201618

[3] http://www.image-net.org

[4] Artificial Intelligence - The Future of Humankind, TIME special issue, 2017

[5] http://www.image-net.org/synset?wnid=n02123159

[6] https://zenodo.org/record/2633630#.XPI0k6eQ3y8

[7] https://www.data-futures.org/freizo.html

[8] https://invenio-software.org

[9] https://hasdai.org

[10] https://zenodo.org

[11] http://projectmirador.org

[12] https://universalviewer.io

[13] https://opencv.org/about/

[14] https://zenodo.org/record/3228296#.XOfmES2Q3RY

[15] https://dtu-wind-energy.freizo.org

[16] https://orcid.org

[17] https://data.dtu.dk

[18] https://scicore.unibas.ch/about-scicore/

[19] https://www.neweuropeanwindatlas.eu

[20] http://www.winddata.com

[21] https://mattermost.windenergy.dtu.dk