INODE

1st Dissemination & User Training Report

Document Due Date: 30/04/2021
Document Submission Date: 30/04/2021

Work Package 10

Type: Report
Document Dissemination Level: Public

INODE
Intelligent Open Data Exploration
is funded by the Horizon 2020 Framework Programme of the EU for Research and Innovation.
Grant Agreement number: 863410— INODE — H2020-EU.1.4.1.3.

(This page has been intentionally left blank)

Executive Summary

This document, INODE Deliverable 10.1, is the first report on dissemination and training activities carried out by the **In**telligent **O**pen **D**ata **E**xploration (INODE) consortium, during the period of project months M1 - M16.

Our mission is to develop a classic unified, comprehensive platform that provides extensive access to open datasets through natural language queries in the fields of Cancer Biomarker Research, Research and Innovation Policy Making and Astrophysics, for a wide range of users from larger scientific communities to the general public.

In this report, we briefly state the guidelines on the dissemination and training activities as proposed before the start of the project. This is followed by an overview on the dissemination and training activities that were accomplished during the course of the project until project month M16, along with plausible future plans in promoting our project.

Project Information

Project Name	Intelligent Open Data Exploration
Project Acronym	INODE
Project Coordinator	Zurich University of Applied Sciences (ZHAW), CH
Project Funded by	European Commission
Under the Programme	H2020-EU.1.4.1.3. - Development, deployment and operation of ICT-based e-infrastructures
Call	H2020-INFRAEOSC-2019-1
Topic	INFRAEOSC-02-2019 - Prototyping new innovative services
Funding Instrument	Research and Innovation action
Grant Agreement No.	863410

Document Information

Authors(s)	Koutrika Georgia, Eleftheraki Stavroula, Glenis Apostolis, Mandamadiotis Antonis (ATHENA) Amer-Yahia Sihem, Patil Yogendra, Personnaz Aurélien (CNRS) Lücke-Tieke Hendrik, May Thorsten (Fraunhofer) Litke Antonis, Papadakis Nikolaos, Papadopoulos Dimitris (Infili) Fabricius Maximilian, Subramanian Srividya (MPE) Bastian Frederic, Mendes de Farias Tarcisio, Stockinger Heinz (SIB) Massucci Francesco, Multari Francesco, Rull Guillem (SIRIS) Calvanese Diego, Lanti Davide, Mosca Alesandro, Guohui Xiao (UNIBZ) Braschler Martin, Brunner Ursin, Kosten Catherine, Smith Ellery, Stockinger Kurt (ZHAW)

Table of Contents

1 Introduction

The aim of this deliverable is to consolidate the impact of the project by spreading knowledge of technical and scientific results among the relevant communities, as well as among the general public. This document was developed within the framework of WP10 and is based on the dissemination and communication plans drafted in our project proposal.

The objective of this deliverable is to provide an insight into the overall objectives, specific strategies for targeting a wide audience, materials and channels implemented for effective dissemination within the consortium and the community, as well as dissemination plans for the near future.

The document is structured as follows. Following a brief introduction in this section, Section 2 describes the dissemination and communication strategy as stated in the project proposal. In Section 3, dissemination materials and services such as the project logo and website are presented. Dissemination and communication activities such as workshops and symposium organised by INODE consortium are described in Section 4.

2 Dissemination and Training Strategy

A summary of the proposed and accomplished dissemination and training activities is given below:

Activity	Proposed	Accomplished till M16/proposed for near future
Organisation of workshops: co-located with major conferences in the field of data science and INODE use cases.	Workshops organized (=2); Participants in each workshop ~ 50 - 100	Round table conference: 1 Workshop: 1 Mini-symposium: 1
Internal trainings within INODE consortium; On-site demonstrations	≥ 2 demonstrations	Internal trainings: 2 Demos: 4
Scientific papers targeting workshops, conferences and journals.	Workshop papers (1-3 per-year) Conference papers (1-2 per-year) Journal papers (1-2 per-year)	Papers: >20
Social networks posts		Blogs: 6 Technologies: 4 Tutorials: 2
Participation in exhibitions and trade fairs showcasing ICT solutions		Showcase INODE a event of Swiss National Science Foundation on Big Data Research

Participation in media (TV , newspapers, radio) events and online workshops, conferences etc., to communicate INODE results		Keynote speeches: 5

3 Training, Dissemination Material and Services

3.1 INODE Logo

A logo has been designed for the INODE project, which is featured on the website and other INODE documents. It is also displayed in all the promotional materials (online or offline).

3.2 INODE Website

The INODE website (http://www.inode-project.eu) remains the prime face of the project. It was developed and has been periodically updated and maintained within WP10. The website has been designed to provide a complete overview of the project's concepts and objectives and to showcase the methodologies and outcomes.

The website has a public area for general audiences, where project-related information relevant for the general public is being disseminated. The "News and Updates" section plays a key role in the dynamization of the site. In this section, news on the advancement of the project, events organized by the INODE team and the events in which partners participated, press releases etc., are communicated. Additionally, a dedicated blog section is available, where the partners come together and publish articles and posts on the different aspects of the project.

Technical information relevant for scientific audiences, such as the description of software components that are included in INODE, are published under "Resources" (Technologies and Tutorials).

3.3 INODE in Social Media

INODE aims to successfully communicate and interact with target audiences through a strong social media presence. For this purpose, a Linkedin profile was set up and is available at https://www.linkedin.com/in/project-inode/. The Linkedin profile is intended to mirror the INODE website. The most relevant blog posts and news from the website are posted in the Linkedin profile.

3.4. INODE Internal Trainings

3.4.1. Astrophysics in INODE: Our Colorful Universe

The MPE members, who provide the astrophysics use case, presented an internal training session for the INODE team members on 08.10.2020. The main focus of this training was to provide an insight into our colorful universe and the SDSS database.

Objects which emit a bluer light are generally hotter than objects emitting a primarily redder light. For example, a star which has a surface temperature of 30,000 K appears blue, while a star that has a surface temperature of only about 3,500 K appears red to the human eye. Our sun is a yellow star with a surface temperature of about 5,600 K. Thus colors can provide insights into the emitting object. But in general, astronomical objects like stars emit light in a range of wavelengths, called the Electromagnetic (EM) spectrum, including a visible region (comprising colors from red to blue) where human eyes are sensitive to and non-visible wavelengths, namely Radio, Infrared, Ultraviolet, X ray, Gamma ray wavelengths.

Astronomers use telescopes to observe the celestial objects and measure the flux by collecting the light emitted through different filters, called photometry or imaging. Since Ptolemy in the second century AD, magnitudes have been used to quantify object brightnesses. The colors of objects are also given in magnitudes corresponding to the ratios of the fluxes in logarithmic scale.

SDSS observes the sky through five different color filters, that is to say, in five different color bands. There are two filters, green (g) and red (r), which fall within the visible region of the EM spectrum. The other three filters correspond to light which is invisible to the human eye, ultraviolet (u), and two infrared wavelengths (i and z). Other colors are designated as u-g, g-r, r-i etc. A star's color can give clues to the important properties of a star, for example, its average temperature. For deriving the temperature of a star, astronomers use a "color-color diagram". Color-color diagrams are generally plotted between two colors for one or many sources.

Another way of studying astronomical objects is spectroscopy. Instead of measuring the integrated light over a wide range of wavelength bands (as in photometry), the radiation from stars, galaxies and other objects are passed through a dispersion medium, for instance a prism, in order to measure light as a function of wavelengths. Spectra can provide us valuable and comparatively accurate insight into the properties of emitting sources like temperatures, elemental abundances, emission mechanisms etc. Stars, galaxies, and quasars are selected from the photometric data for spectroscopic observation. All photometric and spectroscopic data observed over the last 20 years are publicly available online for every one, from astronomers to the general public, for science education.

Interesting science begins when new types of objects that occupy unusual places in the parameter spaces are found. Discovering them, however, is not easy and requires experience and, to some degree, luck, but most importantly, the right sets of tools to easily and intuitively navigate through large datasets.

This is exactly the goal that INODE aims to achieve. Scientists will, for example, be able to input a set of interesting example galaxies into the tools developed by the INODE project. These tools are able to then automatically and intelligently suggest larger samples of similar looking galaxies. The scientists can then increase, refine, or reshape the parameter ranges of interest to arrive at a final set of objects that are of interest to their particular scientific goals. Importantly, since queries can be asked in natural language, this kind of scientific exploration is no longer exclusive to absolute experts in the field. Even amatuer astronomers or simply enthusiastic members of the general public will be able to intuitively navigate through the vast discovery space.

3.4.2. Cancer Research in INODE: Analyzing Biomarkers

The SIB members gave a tutorial on cancer research mostly focused on the OncoMX datasets to the INODE team. The objective of this tutorial was to better inform INODE project members about the cancer research domain, the OncoMX datasets, the defined database schemas (relational and graph-based), and the related ontologies. By doing so, we enabled a better understanding of what the INODE project can contribute to this domain. In the next paragraphs, we summarise this tutorial.

What is cancer? A simplified definition of cancer is "a malignant tissue growth resulting from uncontrolled cell proliferation". Microscopic cancers frequently grow in the human body, but most of the time the immune system reacts and eliminates it. When the immune system is not able to eliminate the abnormal proliferation, there is development of a form of cancer, for example, liver cancer.

What causes cancer? In 1950, the WHO started to realize that migrants developed types of cancer common to their adopted countries: cancers were more often caused by exposures in the environment, rather than inherited genetic factors. In 1965, the International Agency for Research on Cancer (IARC) was created to investigate the causes of cancer in humans. Among the cancer causes discovered over time, we can mention:

- occupational and pharmaceutical agents (e.g., tobacco)
- infectious agents, notably viruses
- natural factors and non-viral infectious agents (e.g., ultraviolet radiation)

Cancer disease mutation. A mutation is an alteration of the DNA (deoxyribonucleic acid), that can cause or predispose an individual to a specific cancer disease. DNA molecules contain genes, that are regions that encode information on how to produce proteins. Proteins are macromolecules, consisting of one or more long chains of amino acids. They are responsible for carrying vital tasks for living organisms, such as pigmentation of skin, transport of elements in the organism, and immune defense against infectious agents. While a mutation is in most cases benign, or repaired by the cell machinery, in some rare cases this alteration can have an impact on the gene products, resulting in malfunctioning proteins, or the incorrect activation of genes. These malfunctioning proteins or inaccurate activations can cause uncontrolled cell proliferation, in some cases leading to cancer.

Cancer biomarkers. To characterize a cancer type, and identify a specific therapy strategy, biomarkers are used. A biomarker is a measurable characteristic used as an indicator of a biological state. Thanks to the use of biomarkers, we have recently achieved advances in cancer therapy and personalized medicine.

An example is the case of Triple-negative breast cancer (TNBC). In TNBC, there is poor response to classical therapies, and an aggressive behavior of the tumor (early relapse, metastatic spread, poorer survival). TNBC abnormal cells present a lack of activation of Estrogen Receptor (ER), Progesterone Receptor (PR), and Human Epidermal Growth Factor Receptor 2 (HER2) genes. TNBC is more common among specific ethnicities, such as Latin, African and African American women, and accounts for approximately 10%–15% of all breast cancer cases. TNBC can be identified with the use of multiple biomarkers, i.e. a biomarker panel. This is because most of the time one single biomarker is not specific to one cancer type, or to one drug response prognosis. For the TNBC example, a biomarker panel could be composed as follows:

- Biomarker 1, no estrogen receptor in cancer cells
- Biomarker 2, no progesterone receptor in cancer cells
- Biomarker 3, decrease expression of Human Epidermal Growth Factor 2

Other biomarkers can be used to identify subtypes of TNBC, such as an increased activation of the EGFR gene.

This biomarker panel implies low disease free survival, and, when the EGFR activation increases, possible treatments with erlotinib, gefitinib, or afatinib drugs.

INODE added value. Providing a question answer (QA) system over cancer research data focused on cancer biomarkers. This system is being designed to be used by worldwide researchers in order to facilitate cancer knowledge retrieval. INODE will enable cancer researchers to query datasets in natural language and to explore cancer-related datasets such as healthy gene expression, cancer differential gene expression, cancer disease mutation and cancer biomarkers. The OncoMX team has provided us with these datasets.

3.5 Collaboration with H2020 Projects and Others

We have meetings with the *NEANIAS project* team to explore opportunities for cross-project technology transfer and other collaboration opportunities. Furthermore, the planning for a joint workshop in the *EOSC Symposium*[1], the annual event for the EOSC ecosystem, is being discussed among several projects (NEANIAS, Dear Cos4Cloud, INODE, TRIPLE, ESCAPE).

In addition, we have a strong collaboration with the *Bio-SODA-project* which is funded by the Swiss National Science Foundation. We collaboratively develop the Bio-SODA systems as well as dissemination and training activities together with the *Bio-SODA-project*.

3.6 INODE-Related Publications

During the course of the project, until M16, our team members have published the following list of articles. Note that all publications are *peer reviewed*. Moreover, in Computer Science, publications in conferences are considered as prestigious as in journals.

Journal Publications

1. Bastian, F. B., Roux, J., Niknejad, A., Comte, A., Fonseca Costa, S. S., De Farias, T. M., ... & Robinson-Rechavi, M. (2021). The Bgee suite: integrated curated expression atlas and comparative transcriptomics in animals. Nucleic Acids Research, 49(D1), D831-D847.

[1] https://www.eoscsecretariat.eu/events/eosc-symposium-2021

2. Ding, L., Xiao, G., Calvanese, D., & Meng, L. (2020). A Framework Uniting Ontology-Based Geodata Integration and Geovisual Analytics. ISPRS International Journal of Geo-Information, 9(8), 474.
3. Papadopoulos, D., Papadakis, N., & Litke, A. (2020). A methodology for open information extraction and representation from large scientific corpora: the CORD-19 data exploration use case. Applied Sciences, 10(16), 5630.
4. Fabian Colque Zegarra, Juan C. Carbajal Ipenza, Behrooz Omidvar-Tehrani, Viviane P. Moreira, Sihem Amer-Yahia, João Luiz Dihl Comba: Visual exploration of rating datasets and user groups. Future Gener. Comput. Syst. 105: 547-561 (2020)
5. Behrooz Omidvar-Tehrani, Sihem Amer-Yahia: User Group Analytics Survey and Research Opportunities. IEEE Trans. Knowl. Data Eng. 32(10): 2040-2059 (2020)

Conference Publications

1. Brunner, U., & Stockinger, K. (2021). ValueNet: A Neural Text-to-SQL Architecture Incorporating Values. IEEE Proceedings of the International Conference on Data Engineering (ICDE), Chania, Greece, 19-22 April 2021.
2. Calvanese, D., Avigdor Gal, Naor Haba, Davide Lanti, Marco Montali, Alessandro Mosca and Roee Shraga (2021). ADaMaP: Automatic Alignment of Data Sources using Mapping Patterns. Accepted in: 33rd International Conference on Advanced Information Systems Engineering
3. Calvanese, D., Corman, J., Lanti, D., & Razniewski, S. (2020). Counting query answers over a DL-Lite knowledge base. In Proc. of the 29th Int. Joint Conf. on Artificial Intelligence (IJCAI). IJCAI Org.
4. Deriu, J. M., Mlynchyk, K., Schläpfer, P., Rodrigo, A., von Grünigen, D., Kaiser, N., Stockinger K.,... & Cieliebak, M. (2020). A methodology for creating question answering corpora using inverse data annotation. In ACL 2020, Virtual, 5-10 July 2020(pp. 897-911). Association for Computational Linguistics.
5. Esfandiari, M., Ria Mae Borromeo, Sepideh Nikookar, Paras Sakharkar, Sihem Amer-Yahia, Senjuti Basu Roy: Multi-Session Diversity to Improve User Satisfaction in Web Applications. TWC 2021
6. O. Gkini, T. Belmpas, G. Koutrika, Y. Ioannidis (2021). An In-Depth Benchmarking of Text-to-SQL Systems. To appear in proceedings of ACM SIGMOD 2021.
7. Seleznova, M., Omidvar-Tehrani, B., Amer-Yahia, S., & Simon, E. (2020). Guided exploration of user groups. Proceedings of the VLDB Endowment, 13(9), 1469-1482.
8. Xiao, G., Lanti, D., Kontchakov, R., Komla-Ebri, S., Güzel-Kalaycı, E., Ding, L., ... & Botoeva, E. (2020, November). The virtual knowledge graph system Ontop. In International Semantic Web Conference (pp. 259-277). Springer, Cham.
9. B. Youngmann, S. Amer-Yahia, T. Milo (2021). Exploring Subjective Databases. To appear in proceedings of ACM SIGMOD 2021.

Workshop Publications

1. Amer-Yahia, S., Anh Tho Le, Eric Simon: Data Pipelines for Personalized Exploration of Rated Datasets. BIAS 2020: 72-78
2. Calvanese, D., Corman, J., Lanti, D., & Razniewski, S. (2020). Rewriting Count Queries over DL-Lite TBoxes with Number Restrictions. In the *33rd International Workshop on Description Logics*. ceur-ws. org.
3. Calvanese, D., Gal, A., Lanti, D., Montali, M., Mosca, A., & Shraga, R. Mapping Patterns for Virtual Knowledge Graphs (A Report on Ongoing Research). In the *33rd International Workshop on Description Logics*. ceur-ws. org.
4. Personnaz, A., S. Amer-Yahia, L. Berti-Equille, S. Subramanian, M. Fabricius (2021). Balancing Familiarity and Curiosity in Data Exploration withDeep Reinforcement Learning. To appear in the Fourth International Workshop on Exploiting Artificial Intelligence Techniques for Data Management (aiDM) in conjunction with ACM SIGMOD.
5. Xiao, G., Lanti, D., Kontchakov, R., Komla-Ebri, S., Güzel-Kalaycı, E., Ding, L., ... & Botoeva, E. The Virtual Knowledge Graph System Ontop (Extended Abstract). In the *33rd International Workshop on Description Logics*. ceur-ws. org.

Demonstrations

1. T. Belmpas, O. Gkini, G. Koutrika. Analysis of Database Search Systems with THOR. ACM SIGMOD 2020
2. A. Glenis, G. Koutrika. PyExplore: Query Recommendations for Data Exploration without Query Logs. ACM SIGMOD 2021
3. Chibah, A., Sihem Amer-Yahia, Laure Berti-Equille: QeNoBi: A System for QuErying and miNing BehavIoral Patterns. IEEE Proceedings of the International Conference on Data Engineering (ICDE), Chania, Greece, 19-22 April 2021.
4. B. Youngmann, S. Amer-Yahia, T. Milo: SubDEx: Exploring Ratings in Subjective Databases. IEEE Proceedings of the International Conference on Data Engineering (ICDE), Chania, Greece, 19-22 April 2021.

Tutorials

1. G. Katsogiannis-Meimarakis, G. Koutrika. A Deep Dive into Deep Learning Approaches for Text-to-SQL Systems. ACM SIGMOD, 2021, to appear
2. G. Katsogiannis-Meimarakis, G. Koutrika. Deep Learning Approaches for Text-to-SQL Systems. EDBT, 24th International Conference on Extending Database Technology, 2021

Keynote speeches

1. “INODE- Intelligence Open Data Exploration”, G. Koutrika, SNTA 2021@HPDC, June 2021

2. "The Rise of Intelligent Data Assistants", G. Koutrika, ACM Distinguished Speaker, ACM-W NITK, April 9, 2021
3. "The Rise of Intelligent Data Assistants", G. Koutrika, Trustworthy Data Science and AI Webinar Series, Simon Fraser University, April 8, 2021
4. "The Rise of Intelligent Data Assistants: Democratizing Data Access", G. Koutrika, BigVis2021@EDBT, March 23, 2021.
5. “Democratizing Data Access Through Intelligent Data Exploration Tools”, G. Koutrika, ACM WomENcourage 2020.

4 Training and Dissemination Outreach Activities

4.1 VLDB 2020 Round Table on “Intelligent Data Exploration”

Kurt Stockinger and Georgia (Yuli) Koutrika initiated a Round Table along with the speakers Sihem Amer-Yahia, Jignesh Patel, Carsten Binnig, João Pedro Monteiro as part of the 46th International Conference on Very Large Databases (VLDB 2020).

In this round table, the participants discussed “Data exploration” extensively, with a focus on what it takes to bridge the gap between users and data, and the new generation of intelligent data exploration tools that are emerging at the intersection of data management, natural language processing, machine learning and visualization.

4.2 INODE/EOSC Workshop

This workshop will take place on May 21st, 2021. The workshop aims at bringing EOSC members an insight into the INODE system. Integration and exploitation of the INODE services will be promoted by EOSC-hub. The INODE system will eventually be integrated with EOSC-hub services and made available for a large international group of EOSC users.

The workshop provides a comprehensive view on where INODE stands and how the INODE consortium is preparing for the challenges ahead, until the completion of the INODE project in October 2022. Short presentations from the INODE use case providers are intertwined with short demos from technical teams.

https://eosc-portal.eu/events/inode-project-eosc-workshop

4.3 Mini-symposium at the Platform for Advanced Scientific Computing (PASC) 2021

As part of a joint activity between INODE and the Bio-SODA project, founded by the Swiss National Science Foundation, this year we are organizing a minisymposium that was accepted at the PASC21[2] on July 5th to 8th, 2021. This minisymposium brings together international speakers that work on semantic data integration and interoperability in life sciences. We will discuss how to combine dispersed data sources such as web pages, databases, and knowledge graphs as a continuum. Another subject to discuss is about the swift integration of knowledge in Wikidata through linked data (RDF) and schemas (Shape

[2] https://pasc21.pasc-conference.org/program/minisymposia/

Expressions, ShEx)[3]. Moreover, we will talk about an approach for semantic integration in metabolism with an automated chemical ontology expansion. Finally, we will present our work on triple extraction from cancer biomarker literature with INODE as well. Further information about this minisymposium is depicted below:

Title: Toward Semantic Integration of Biological Resources

Abstract. One major potential and promise of big data analysis lies in the simultaneous mining and integration of multiple heterogeneous sources of data. In life sciences, recent years have seen the increasing availability of biological and bioinformatic databases using the Resource Description Framework (RDF), which facilitates automatic data processing and interoperability. However, there are major stumbling blocks on the path to mass adoption. The complexity of general-purpose models, inconsistent data models, and low usability are some of the challenges that hamper the use of RDF resources by the bulk of biological researchers. This mini-symposium brings together specialists on semantic data integration in life science and will provide a forum to explore innovative solutions to fulfil the potential of big data integration.

Organizers: Tarcisio Mendes de Farias (Swiss Institute of Bioinformatics, and Kurt Stockinger (Zurich University of Applied Sciences), and Christophe Dessimoz (University of Lausanne, University College London).

Domain: Life Sciences

Invited speakers: Dr. Janna Hastings (University College London, UK), MSc. Dimitris Papadopoulos (INFILI technologies company, Greece), Dr. Franck Michel (University Côte d'Azur, Inria, CNRS, I3S (UMR 7271), Biot, France), and Dr. Andra Waagmeester (Micelio software development company in Antwerp, Belgium).

[3] https://shex.io

5 Conclusion

This deliverable provides a brief overview of all the dissemination activities and training carried out by the INODE team during the course of the project, till M16.

The INODE team has focused on raising awareness among relevant communities by organizing online events and participation in national/international online events. Dissemination and training activities have unfortunately had some setbacks due to COVID restrictions, since the beginning of the project. Thus the proposed in person events have been postponed.

Overall, partners have published many peer reviewed articles (a few are currently under the review process) in key scientific meetings and conferences. INODE team members have been successful not only in developing full-fledged intelligent data exploration systems, but also in spreading awareness and communicating results to a wide range of audiences from the general public to experts in the relevant communities.