Open Call Collection OC-2018-1

Proposal Reference OC-2018-1-22782

Title: **design, Analysis, Management of data lakes In a Cost Action**

Acronym: **AMICA**

Summary

The large number of data available in open data repositories, corporate datasets and documents constitutes an extremely large, heterogeneous and evolving information space. Modern organizations and individuals have realized the value of Big Data and have started to collect massive amounts of data from almost every aspect of their activities. The goal is to later analyse it and turn it into useful insights that can help decision making, and improve their business. At the time of collection, it is often not clear how the data will be used. Thus, it is stored in its native form. This results in datasets of high volume and heterogeneity, making traditional data management systems inadequate in keeping that data. A new architecture of repositories has started to become prevalent, the data lake. Data lakes are still at the beginning and open research challenges have to be addressed to make these systems efficient and effective.

The aim of this Action is to elevate data lakes into modern data management systems by studying a) possible usage scenarios, users' profiles, and functionalities required by the users; b) a reference functional architecture and its main building blocks; c) techniques for summarizing the contents of the data sources; d) techniques for linking, discovering, exploring, analyzing, and merging the sources based on the summaries.

To build a successful solution, due to the multifaceted nature of the problem, the Action proposes to create synergies between all possible interested actors: data owners, technology providers, computer science researchers, and decision makers.

Key Expertise needed for evaluation

Electrical engineering, electronic engineering, Information engineering
Databases, data mining, data curation, computational modelling

Keywords

data lake

big data

data analytics

data management

data exploration

Funded by the Horizon 2020 Framework Programme
of the European Union

TECHNICAL ANNEX

1. S&T EXCELLENCE

1.1. CHALLENGE

1.1.1. DESCRIPTION OF THE CHALLENGE (MAIN AIM)

We live in an era of ubiquitous data, where organizations and individuals produce and publish data at unprecedented rates, generating an extremely large, highly heterogeneous and rapidly evolving information space. The Public Sector is largely contributing to create and keep updated this large number of sources, in particular in Europe, where community directives and national legislations promote the use and the reuse of data. This enables a desirable cycle in which data generated by the Public Sector can be used as raw material for innovative value-added services and products which boost the economy by creating new jobs and encouraging investment in data-driven sectors (Commission notice 2014/C 240/01). The socio-economic impact generated by the available data assumes a paramount importance. The European Data Market Study SMART 2013/0063 (available at: http://datalandscape.eu/) demonstrates that the value of the data economy will increase to EUR 739 billion by 2020, representing 4% of overall EU GDP (more than doubling the situation today), and the number of data professionals will increase from over 6 million in 2016 to over 10 million by 2020.

Traditionally, enterprises collect and analyse data of interest by means of data warehouses. Data warehouses are designed to address pre-defined analytic needs, and populated via extraction, transformation and loading (ETL) processes involving specified data sources. Therefore, such applications are not flexible: a change in the sources or the need for a new kind of data analysis requires the adaptation of the warehouse schema and the ETL processes to address the new request, with a possible costly impact on the entire framework. This scenario has changed significantly in the recent years. Big Data has illustrated a huge potential in discovering new facts the existence of which was practically unknown. Platforms supporting Big Data management and analytics have been developed, most of them revolving around distributed file systems and MapReduce implementations of processes for managing and analysing the data.

However, we are witnessing a new phenomenon. Modern organizations are collecting data from almost every aspect of their daily activities to analyse it and turn it into valuable insights that can help in taking informed decisions and driving their business. In these scenarios, the Big Data cannot be thought as a large monolithic data source, but is formed of a plethora of heterogeneous, in format, dimension and content, data sources. Big Data is the result of the unprecedented race in data collection that organizations are doing. Their claim is that the data has to be collected, as much as possible, in its native form because we cannot know in which extent and for which purposes it will become useful for the future for the organization. Data warehouses and Big Data platforms have so far demonstrated they are limited for this new wave of data, replicating the same issues that have plagued the previous architectures: the creation of data silos, where the data is stored for addressing a specific task and the inability to deal with the data heterogeneity (e.g., streaming and unstructured data). A new form of architectures has been born to take up that role: the data lake.

A data lake is a storage space for large amounts of data of different kinds. Data lakes are different from data warehouses. The kind of data contained in a data warehouse has been carefully decided upon in advance by a data engineer; instead in a data lake, any kind of data can be placed, i.e., structured, semi-structured, or unstructured (text data). Furthermore, any data that needs to be placed in a warehouse will have to be first transformed to fit a specific schema, often leading to some information

Funded by the Horizon 2020 Framework Programme
of the European Union

loss. At the rate data is produced nowadays there is no space for such pre-processing, and data is stored raw in a data lake. Furthermore, data warehouses are based on database technologies that do not scale well. Instead, data lakes operate under the assumption that the space is cheap and use distributed file systems and store information in its most native format. Last but not least, data warehouses are designed for business professionals and serve specific functionalities, while data lakes are built for data scientists, and are thus generic.

The task of a data scientist (and other IT professionals) is to explore the vast amount of data that is available in the data lake, identify those parts that are of interest, select what best fits a specific task and proceed with processing and analysis. Unfortunately, very little effort has been invested in developing high-end tools and services that could support and facilitate data scientists in the above tasks. Data scientists have to manually search or write from scratch any scripts needed to identify, retrieve, and transform the elements of interest for analysis. As data lakes grow larger and larger in terms of number of datasets, heterogeneity, and size, data exploitation is becoming laborious, time consuming and error prone. Providing the right tools to support data scientists in searching the data lake datasets is of paramount importance. It can lead to better analytic results, more trustful insights, more fine-grain analyses, optimized data management, and shorter time to decision making.

The motivation of the Action is the fact that (a) the research community has put a lot of effort into supporting the availability and interoperability among sources through techniques for publishing, semantically enriching, matching and integrating data sources; (b) policy and decision makers have invested a lot of effort in supporting a “data-driven culture”, promoting open publication, through the use and the re-use of data; (c) public and private organizations have broadly accepted this challenge by publishing large amounts of high quality data on the web; and (d) tools and applications for the automatic discovery, use and re-use of such data sources are still limited in their capabilities of discovering and analysing these data.

The vision inspiring this Action is to recognize datasets as a significant and vital part of a data lake and to study the technology and software ecosystem around data lakes. To do so, the Action draws its strategy across three fundamental pillars: (i) **study of techniques for designing data lakes**, through the collection of real world scenarios where the goal is to discover the analytics operations and the functionalities required in data lakes, and to experiment with the technological solutions proposed; (ii) **study of techniques for building data lakes**, where the challenge is the effective dataset modeling through summarized descriptions supporting/based on flexible indexing and linking/matching across datasets; (iii) **study of techniques for using and managing the data lakes**, where the Action will investigate the efficient and effective dataset discovery through novel models of query answering on profiles and content; dataset combinations as answers; and dataset quality-based selection for a task at hand.

To build a successful, unified and effective solution - due to the multifaceted nature of the problem – the Action proposes to create synergies between several disciplines, such as Big Data Management, Semantics, Semantic Web, Information Retrieval, Artificial Intelligence, Machine Learning, User Interaction, and Natural Language Processing.

1.1.2. RELEVANCE AND TIMELINESS

1. A recent report published by the European Data Portal [ODM16] highlights that “by 2016, the majority of the EU28+ countries have successfully developed a basic approach to address Open Data.” and “the market size of Open Data is expected to increase by 36.9% to a value of 75.7 billion EUR in 2020”. Open Data represents an important asset for the (European) economy. The Action will develop techniques for managing and analysing open data, thus allowing companies and the public sector to exploit this data for their businesses. The high amount of data, of various degree of quality, has to be associated with effective and efficient means to exploit it.

2.Open data is only one of the possible “ingredient” for lakes. Forbes [F17] reports that the annual data creation rate is 16.3ZB. To be effective in real business scenarios, data lakes may need to manage both open and proprietary data (including confidential company information and private personal data).

3. Another report, [CFS15], investigates the skills required to manage and glean insight form data. The outcome of the study is that the matter requires hard skills (the ones typically owned by IT Professionals) and soft skills (e.g., collaboration, problem solving and communication, curiosity, creativity). The Action will allow researchers and other stakeholders with different skills to meet together and to promote together research activity on the Action topic.

4. Big Data research is mature enough to provide advanced analytical solutions. From a research perspective, techniques and tools for identifying, cleaning, integrating and wrangling data have been developed [KKA+17]. Furthermore, approaches for data ingestion, transforming unstructured and semi-structured data into structured data (e.g., RDF triples [FRI+15]), have also been proposed, and data provenance has been applied to the field [SP16]. A number of projects have contributed to the definition of a “substrate” of research outcomes (see Section 1.4.2) that will be leveraged by the Action. What is missing is an “engineering process” to identify and standardize the components of the data lake ecosystem, design the functionalities, provide benchmarks for testing the quality, thus turning these approaches from ad-hoc to industrial applications. From the industrial perspective, a number of platforms for data analytics and data integrations have been developed (see for example the Gartner Magic Quadrant for Data Science Platform 2017 reported in [GMQ17]). Nevertheless, the main focus of these platforms is offering tools for the management and analysis of large dimensional datasets. The Action aims at defining a reference functional architecture for this kind of applications, studying a standard set of functionalities, and specifying a complete ecosystem for data lakes, where the problem is not only the large amount of data, but also the large amount of data sources and their heterogeneity.

1.2. OBJECTIVES

1.2.1. RESEARCH COORDINATION OBJECTIVES

The Action is built around three main pillars, with each one introducing a number of research coordination objectives:

1. Designing the data lake. The Action will promote:

Obj 1.1) the classification of the possible users of data lake applications, the definition of the high level functionalities and the analytic processes they need.
Obj 1.2) the collection of datasets and scenarios to be used by researchers and practitioners for their experimentations. Scenarios involving open and private data will be investigated.
Obj 1.3) the specification of a reference architecture for a data lake, defining its main software components in terms of functionalities, inputs, and outputs.

2. Building the data lake. The Action will promote:

Obj 2.1) the definition of a critical analysis of the emergent techniques for profiling data sources, identifying and critically reviewing the most promising approaches and enabling the development of a conceptual map of the open and closed issues related to the research topic.
Obj 2.2) the development of automatic, expressive, complete and effective techniques for dataset profiling (advanced indexing), i.e., the creation of summarized content descriptions for whatever resides in the data lake.
Obj 2.3) the development of automated solutions for data linkage and merging, that will be able to perform not only across dataset values but also across meta-data descriptions, i.e., dataset profiles.

3. Using and managing the data lake. The Action will promote:

Obj 3.1) the definition of a critical analysis of the emergent techniques for discovering and exploiting data sources in a lake, identifying and critically reviewing the most promising approaches and enabling the development of a conceptual map of the open and closed issues related to the research topic.
Obj 3.2) the development of advanced user-friendly query methods and analytic functions that will allow data scientists and other users to search, identify, analyse or more generally exploit the datasets, or the parts of the datasets that are related to some specific goal they have in mind.
Obj 3.3) the development of benchmarks to evaluate and compare the framework developed.
Obj 3.4) the development of advanced techniques, e.g., provenance tracking, trusted versioning control, for evaluating the quality of the datasets in the data lake through an analysis of the profiles.
Obj 3.5) the study of approaches for data-lakes curation, maintenance and (self-)management, through techniques for logging and analysing the usage of data lake, which would improve the capability of the lake to address the users’ needs. Autonomic computing principles can also be used for this including self-awareness, self-protection, self-configuration, etc., and utilizing for that recent advances in Artificial Intelligence (distributed AI, multi-agent systems, machine (deep) learning, etc.). This Objective also includes the study of all the issues related to the data governance and the needs to merge and support the existing enterprise data management paradigms, tools, and methods. Data protection and issues

related to the General Data Protection Regulation (GDPR), as well as mechanisms that can assist in their enforcement, are also of interest.

1.2.2. CAPACITY-BUILDING OBJECTIVES

The main capacity building objectives of the Action are:

c1) launching and establishing a cooperative, multi-disciplinary network of researchers, practitioners, application domain specialists and possible users (constituting a community of active stakeholders) for promoting the development and the effective use of data lakes for managing, discovering and exploiting large amounts of data sources. The Action will promote an active and transparent participation of the members: via "calls for volunteers" the participants will be directly involved in the Action by requiring their support in the definition and organization of the events (meetings and training school); via "call for participations", the members eligible for reimbursements, the STSM (Short Term Scientific Mission) applications to be funded and the conference grant to be delivered to researchers from ITC (Inclusiveness Target Countries) will be defined. The claim is that the more people are actively involved, the more the Action and the people themselves will benefit from it.

c2) building a critical mass of research activities and outcomes that achieve the sustainability of the research themes beyond the Action through:

a) the collection of the most relevant material (slides, tutorials and other training materials, scientific papers, prototypes, datasets, scenarios and benchmarks) developed within the Action in an open annotated bibliography for improving their dissemination, reuse and extension.
b) the enlargement of the network of academic, professional researchers and other stakeholders beyond the ones initially involved during the preparation of the Action proposal.
c) the development of lectures for Master Thesis and PhD courses, guaranteeing the persistence and circulation of the methods and applications relevant for the Action.
d) the organization of open meetings (at least 3 during the Action), with brainstorming sessions and hackathons, enabling strong cooperation among the participants, and the establishment of new and durable research relationships. The organization of meetings mainly in ITCs will enable the local visibility of the Action and a large participation of local researchers and interested people.

c3) supporting the discussions and idea sharing among early-stage and senior researchers by means of training schools and symposia, and by promoting visiting activities and exchanges through Short-Term Scientific Missions (at least 50 during the Action). Criteria for improving the participation of Early Career Investigators (ECIs) and researchers from ITCs will be implemented.

c4) facilitating the transfer of knowledge and technology to the scientific community, practitioners, enterprises, and other stakeholders, through:

a) the dissemination of the Action results through high quality conferences and journals;
b) the development of targeted publications, newsletters and discussion groups;
c) communications exploiting the capabilities offered by the most common social networks (e.g. by means of the creation of a Twitter channel, Facebook Page, LinkedIn group, YouTube channel, etc.)
d) the participation and organization of meetings and events with people from industry or government promoting the reuse of the Action outcomes in professional applications;
e) the development of scenarios, use cases, benchmarks, prototypes and libraries of source code;
f) the definition of project proposals to national and international (H2020, FP9) calls for applications.

1.3. PROGRESS BEYOND THE STATE-OF-THE-ART AND INNOVATION POTENTIAL

1.3.1. DESCRIPTION OF THE STATE-OF-THE-ART

Data Lakes. A data lake is a central location where to store the enterprise data regardless of its sources or formats [LS16]. This architecture introduces a large flexibility: the lakes are fed information in its native form and the data administrator does not have to design a data structure shared with all the sources of interest. However, this flexibility comes with a price: the retrieval and reuse of data sources in the lake can require the development of ad-hoc applications implementing complex processes. Different tools are needed for implementing lakes. Usually organizations use Apache Pig (https://pig.apache.org/), Hive (https://hive.apache.org/), Spark (https://spark.apache.org/) and other applications implemented with the MapReduce programming model. Some approaches have been proposed in the literature, in particular as software demonstrations of specific functionalities [HCQ16, MT17].

Creating, managing, using summarized descriptions of data sources. Data profiling, i.e., the task of creating small informative summaries of databases [J09], is an activity of paramount importance for any IT professional. Typical scenarios mainly include (big) data exploration, providing users with an insight of a data source content, data analytics, where profiles allow the detection of outliers [WM13] and the interpretation of the results [LBL16], data integration, supporting the identification of sources to integrate, data management (data indexing, compression and query management). The automatic creation of the data profiles is a challenging task, due to issues related to the data volume (e.g., the number of integrity constraints usually grows quadratically - or more - with the schema size), the variety (e.g., data profiles summarize the sources under different perspectives) and the integration of the profiles (e.g., we can have a full insight of a dataset only combining a different number of profiles describing it) have to be addressed [KHW+17]. Profiles typically describe statistical distribution of single, or multiple columns, and dependencies between columns [AGN15, KHW+17, DFA+17]. Only a few tools have been proposed for the management and the automatic combination of profiles in solving tasks. Among them, [FAM+16] has identified (but not implemented) an initial set of primitives, classified in schema-driven, data driven and enrichment; data civilizer [DFA+17] and KAYAK [MT17] implement some basic functionalities for storing and retrieving profiles, finding similar datasets, discovering sources about a specified topic.

Querying data. Traditional query answering methods of structured queries for keyword search on relational data which are usually classified in schema-based or graph-based [YQC10] approaches. Schema-based techniques exploit the schema information to issue SQL queries with the same meaning as the original keyword queries. On the other hand, graph-based techniques treat relational databases as graphs. Solving keyword queries in this context requires the computation of specific structures over the graphs (e.g., Steiner trees). Approaches exploiting graph-based representation of the data source content have also proposed for querying RDF sources [EB11]. A few proposals have adopted information retrieval (IR) style solutions (typically based on vector space models or probabilistic retrieval models [ZM16]), where the main issue is building meaningful and integral information units from the data (which in a relational database is spread over a number of tables) to index and rank according to the given query. In [FLW11], these atomic pieces of data are called tuple units.

Data Quality through data profiles. Existing data quality approaches [CM+15, FGW11, BMNT11] are typically based on irregularities of the data values to provide an assessment on the data contents. The challenge here is dealing with large amount of data sources with diverse degrees of veracity and with their summarized descriptions. Unstructured data crowdsourced by users might be an important source for a Data Lake but their quality should be much lower than the one provided by Open Government portals.

1.3.2. PROGRESS BEYOND THE STATE-OF-THE-ART

In the last few years, research on Big Data management and analysis has become a hot and challenging topic. Only recently has the research community moved from devising approaches for large amount of data into the development of techniques for large numbers of heterogeneous data sources, i.e. the data lakes. Even if a number of approaches and techniques has been developed, there is space for further research in the area. This is a multi-disciplinary work that requires a joint effort from different disciplines. This is the work that an Action can effectively promote. In particular, the Action can support progresses in these areas:

- While the building blocks that can be used to achieve data lake functionality exist, and are available commercially, no study yet exists to determine which building blocks, in which configuration.
- The use of summarized descriptions of the data sources in lakes can make the framework management scalable. There is a rich literature on profiling relational sources with statistical metadata. Nevertheless, the use of other kinds of knowledge to create profiles (semantics, lexical,...), and a deep evaluation of the exploitation of profiles in this area, and the generation of summarized descriptions for other kinds of sources (text, streaming) are still a challenge.

- There is no theoretic / complete framework listing or description of the possible operations on sources in the data lakes and in their summarized descriptions.
- Techniques for discovering sources relevant for satisfying a specific information need or to be analysed for a specific purpose based on descriptions are still under development. Only few approaches and vision papers have been proposed.
- Novel ways to evaluate the quality of the datasets that take into consideration the insight on data source contents provided via the profiles can be useful to assess the quality of the data lake content.

1.3.3. INNOVATION IN TACKLING THE CHALLENGE

The Action will make possible an innovative approach to the challenge starting from and mixing two points of view: a technology centric and a user-driven perspective. This fact represents one of the main benefits that the Action can provide. The problem on hand has a multifaceted nature, due to the heterogeneity of the datasets and the possible operations that users may wish to apply on the data. A complete solution to the challenge requires the exploitation of synergies between several disciplines, an effort that it is typically not achievable in funding frameworks other than COST due to the specific focus that is usually asked to the applications and the numerous areas of expertise required.

A technology centric perspective studies the problem starting from the analysis of the available technology and the features of the data involved. In the Action, the main computer science areas involved are: Big Data management and Analysis, where the proposed techniques can be adapted and extended to work with data lakes; Information Retrieval and NLP, where techniques to summarize unstructured text and retrieve data profiles can be applied; semantic web, that allows structured data to be available on the web; statistics, machine learning and data mining, supporting the profiling, clustering and the analysis of the datasets. Moreover, techniques for analysing open data require the development of spatial and temporal data analysis, where the use of data profiling is a challenge.

A user-driven perspective starts from the user requirements and needs and develop approaches to deal with them. For example, how to enable data lakes to be exploitable both my machines which might use APIs and query languages and humans which should be exposed to visualizations and to natural language interfaces. The Action will involve users and stakeholders with different needs to have a large spectrum of knowledge as to what will be required.

Brainstorming and networking sessions planned in the meetings organized by the Action will allow the intersection of these perspectives and to obtain a complete insight of the main challenges to address. This effort is also needed for the quick development of business applications allowing to bridge the gap between user requirements and available software solutions.

1.4. ADDED VALUE OF NETWORKING

1.4.1. IN RELATION TO THE CHALLENGE

The achievement of the Action Research Coordination objectives and Capacity-building objectives requires the union of knowledge between several computer science disciplines and to mix diversified technical skills. Only through a COST Action it is possible to connect researchers:

- with expertise in different Computer Science areas, ranging from Big Data management, semantics, Semantic Web, information retrieval, artificial intelligence, machine learning, user interaction, service science, service design, and natural language processing.
- with different research skills, ranging from theoretical to actual software development.
- with different goals, including academic researchers aiming to start new research projects and identify new challenges, ECIs with the goal of consolidating their skills, Trainers, who want to find new materials, scenarios for their activities, industrial researchers, practitioners, who see in data lakes a new business opportunity, data owners (including public administrations, and decision makers in general) that aim to be able to use for specific purposes their data assets.
- from different countries, with heterogeneous availability of (open) data, opportunity of participating in research projects in challenging areas.
- with different levels of technological infrastructures available, in an area where the capacity of managing and working with large amount of data is required.

The flexibility of the scheme, that allows researchers to continue the research activities they already have in place, and to schedule specific research goals year-by-year, the autonomy of participants in relation to how they carry out their research activities, the possibility to open the Action to further participants, while it is in progress, the large support for training and exchanges of researchers, the large emphasis on the joint development of new ideas, all make this COST Action the best way to integrate and merge outcomes from different research areas. Moreover, the support for networking activities enables the development of a critical mass of researchers in the topic that may go beyond the Action goals and duration, enabling strong relationships to be built and consequently collaboration opportunities to be created in other research activities derived and related to the Action. The most important aspect is that these relationships can become long-term collaborations beyond the end of the Action, and capitalized by the participant for a large part of their business activity.

1.4.2. IN RELATION TO EXISTING EFFORTS AT EUROPEAN AND/OR INTERNATIONAL LEVEL

The research ideas proposed by the Action have been only partially addressed by other research projects. Nevertheless, there is a large and active community working on big data research and issues related to big data management and analysis have been addressed by several H2020 calls and projects.

Among the communities of researchers working in these fields, the Big Data Value Association (http://www.bdva.eu/) is potentially a privileged interlocutor for the Action. It includes more than 150 organizations working in the field and aims to boost European Big Data Value research, development and innovation and to foster a positive perception of Big Data Value. Another interesting reference is the Big Data Network (www.big-data-network.eu), a network that include 55 centres of excellence that, in Europe, perform cutting edge research and drive the development and evolution of all aspects of Big Data. The Network, that aims at increasing exchange and collaboration in research, identifying synergies for educational collaboration and exchange of data scientists, and its constituting centres, definitely represents an interesting point of reference for the Action enlargement and dissemination. High Performance Computing (HPC) can support the data lake technology. In this area, the main possible references are: 1. PRACE (http://www.prace-ri.eu/), SESAME NET (https://sesamenet.eu/) and ETP4HPC (http://www.etp4hpc.eu/). PRACE is an association that includes 25 member countries whose representative organisations aim to create a pan-European supercomputing infrastructure, providing access to computing and data management resources and services for large-scale scientific and engineering applications at the highest performance level. SESAME NET and ETP4HPC are networks of HPC competence centres aiming to facilitate access from this collection of HPC services.

Moreover, there is a large number of projects and large communities working with open data. This is of interest for the Action since open data (in particular the data available from the Public Sector) can be used for populating and extending the private data sources stored in data lakes. The European Data Portal collects more than 450K open datasets published by the EU Public Sector. The development of public datasets has been the goal of recent calls issues by INEA in the CEF Telecom framework. The first results of these projects will be available in the next months.

Finally, big data management and analysis is still a hot topic for the H2020 programmes. The Action is completely in line with what the EU Commission proposes in most of its research and innovation funding schemas (see for example the H2020 ICT work programmes https://ec.europa.eu/programmes/horizon2020/en/h2020-section/information-and-communication-technologies), where calls for dealing with the storing, management and analysis of large amount of data are proposed. Some calls from the current work programme, in particular, focus on the need of big data management and analytic tools (ICT-12-2018-2020) and on the creation of metadata and metadata standards to achieve a high level of abstraction for describing data and software resources (ICT-16-28). Finally, a large number of calls highlight the impact of big data management and analytics research for better informed decision and policy making. Example of this are calls H2020-SC6-TRANSFORMATIONS-2018-2019-2020 on “Using big data approaches in research and innovation policy making” and H2020-SC6-DT-TRANSFORMATIONS-02-2018-2019-2020 on “Transformative impact of disruptive technologies in public services”.

2. IMPACT

2.1. EXPECTED IMPACT

2.1.1. SHORT-TERM AND LONG-TERM SCIENTIFIC, TECHNOLOGICAL, AND/OR SOCIOECONOMIC IMPACTS

The Action will enable research groups in the participating COST countries, with heterogeneous research expertise, different research methodologies, and working in different Computer Science fields related to the management and analysis of data, to collaborate with each other and to establish long term professional relationships. The main benefit is for ECIs and younger researchers who can improve their research methodologies, learn about the research activities carried out by different research groups, and benefit for a longer period of this networking. The Action will impact at the scientific, technological and economic level as follows.

Short-term Impact

a) The promotion of multidisciplinary network of researchers working in hot and challenging computer science fields involving people from research groups with different expertise and experience levels in different research areas (scientific-technological impact)
b) The identification of the main needs, ranking the main priorities, and promoting the research in a challenging field (scientific-technological impact)
c) Scientific outcomes as research projects, and publications in International Conferences and Journals (scientific-technological impact)
d) The development and the promotion of a reference functional architecture for data lakes including the definition of the main composing building blocks. (scientific-technological impact)
e) Promotion of common research projects involving academia and industry bridging the gap between these organizations and guaranteeing professional opportunities to the researchers (scientific-technological impact)
f) A systematic classification of the existing literature in the area published in the form of an annotated bibliography (scientific-technological impact)
g) The creation of new courses/educational material for graduate level courses from the material collected and published. This will enable new professional position and knowledge transfer (scientific-technological / socio-economics impact)

Long-term Impact

a) The development of software applications implementing / reusing / starting from the research activities carried out in the Action (scientific-technological impact)
b) The network of researchers and the relationships built will continue beyond the end of the Action, and will lead to new research activities, new applications in calls for projects and new visiting activities between researchers (scientific-technological impact)
c) The development of tools easing data exploitation, thus improving the reuse of data, promoting a data-driven culture and the ability to make data-based decisions (socio-economics impact)
d) Scientific innovation, improved training, new research expertise and new skills for the organizations of Action members (scientific-technological / socio-economics impact)
e) The Action develops tools for adding value to Open Data, especially those provided by governments. Citizens will be able to identify and collect the right data related to some topic of interest when searching in open data repositories. This indirectly promotes transparency, democracy and active citizen participation (socio-economics impact)

2.2. MEASURES TO MAXIMISE IMPACT

2.2.1. PLAN FOR INVOLVING THE MOST RELEVANT STAKEHOLDERS

The relevant stakeholders for the Action belong to these categories:

1. Researchers, from Academia and Industry, working in the areas of Big Data Management and Analytics, Semantics, Data Quality, Provenance, Semantic Web, Open data, Natural Language Processing, Machine Learning, Data Mining which will develop techniques for building and using summarizing descriptions of the sources.
2. Technology providers, and vendors, typically from Industry, who will exploit and apply the Action ideas and the outcomes which will result in the development of business applications.
3. Data owners, policy and decision makers, public sector organizations, and governments produce large amount of data and need to manage and analyse data for their business activities.

Even if the Action is oriented to study data management, analysis and usage from the computer science perspective, researchers from the Economics disciplines will be asked to assess the impact of data lakes in exploiting the available data for policy and decision making in public and private scenarios.

Specific strategies involving these stakeholders are planned as part of the Action, including:

- The exploitation of the network of the relationships of the Action Members. Task 1.1. focuses on profiling the possible Action stakeholders with respect to their data lake usage, requirements and problems. The pool of interview candidates for this profiling activity will be expanded via spiralling out to additional contacts generated by each interview.
- Formal calls for participation will be issued and sent to a number of relevant Organizations (both from Public Sector and from Industry).
- Diversified dissemination of the activities promoted by the Action and the scientific outcomes, including flyers, Twitter, a newsletter, a mailing list, a website, tutorials, demonstrations and papers.
- Organization of events such as International Conferences, special issues in top-ranked journals and Training schools in order to promote the Action results to the attendees, inviting external stakeholders as participants and speakers, and organizing hackathons to enlarge and diversify the possible interested participants.
- Participation in events organized by the participants' Institutions (e.g., Open Days, Industrial Days, …) or by external Entities (e.g. Workshops to be co-located in important Conferences).

2.2.2. DISSEMINATION AND/OR EXPLOITATION PLAN

The dissemination plan has two directions: towards the members of the Action (i.e., internal dissemination) and to the scientific community at large in addition to other interested stakeholders and individuals that are not participating to the Action, particularly government and industrial partners.

The internal dissemination is a crucial activity for the success of the Action since it allows members to take advantage of the opportunities offered by the network, e.g., new collaborations in research projects, the availability of technical and training materials, the presence of researchers with complementary skillsets and research needs. The dissemination towards external stakeholders is crucial for enabling usage of the Action as a reference in the research panorama thus improving the attraction of stakeholders in its activity and increasing the opportunity for the participants and the quality of the outcomes. It also increases the immediate utility of the Action for government and industrial partners.

The dissemination / exploitation plan is based on the following components:

- Action website. A technological infrastructure centred on a highly interactive website that will allow members to disseminate news relevant for the Action and external stakeholders to know the main events organized and the outcomes achieved. The website will publish also a) a database of the involved researchers, each one with a personal profile and the ability to publish news / requests for cooperation / research success /…; b) an annotated bibliography, where the users can insert, reviewed and comment on the existing literature; c) all the Action events (meetings, open conferences, training schools, call for participations) to increase the participation in the Action; d) all the training / research material produced by the Action (presentations, results of the brainstorming sessions and hackathons, lecturers, …).
- Targeted publications to reach the possible stakeholders, including flyers, the Twitter feed, a newsletter, mailing lists, and the scientific communication through papers, articles, demos and tutorials.
- The participation and organization of meetings and events with attendants from industry and government promoting the reuse of the Action outcomes in professional applications.
- The development of scenarios, use cases, benchmarks, prototypes and libraries of source code, thus providing to the interested people the flavour of the Action work.
- Publication of a Final Report describing the main results achieved.

2.3. POTENTIAL FOR INNOVATION VERSUS RISK LEVEL

2.3.1. POTENTIAL FOR SCIENTIFIC, TECHNOLOGICAL AND/OR SOCIOECONOMIC INNOVATION BREAKTHROUGHS

Innovation Level. The Action has a huge potential for scientific, technological and socio-economic innovation breakthroughs. Data lakes have been recently proposed as a reference framework for collecting, managing, analysing and using the large amount of data available. This framework fully revolutionizes the traditional way of conceiving data analysis, typically relying on enterprise data warehouses. Data warehouses adopt “schema on write” data management [LS16], which requires an enormous time and cost of preparing the data. The schema is built around pre-defined analytics goals and available data sources to populate it via customized ETL processes. High costs for adapting the approach are required if something changes. Data lakes adopt a “schema on read” approach, where custom schemas are specified by the users into their queries. This is a flexible approach, since data sources are stored in their native formats and used for addressing unspecified users’ information needs. Nevertheless, complex processing and a large effort are required in building the applications that address the users’ need when specified.

From the scientific and technological perspective data lakes have tremendous potential. Nevertheless, a large amount of work is still required since there is no standard reference architecture and the current applications are built around specific scenarios developed around the user requirements. All the existing data lakes are based on parallel computing technologies for managing and analysing data, but differ on the components forming the lake, the functionalities provided and the purposes addressed by the applications. This lack of process engineering has a high cost for software houses that are forced to reinvent the wheel every time they deliver a new lake. The Action can support the data lakes to move from ad-hoc to industrial products. This process requires joint multi-disciplinary skills and expertise which cannot be found in a single research group. The networking activities supported and the ability to join different stakeholders make a COST Action the ideal conduit to perform the process of engineering and standardization of all the data lake design, building and usage processes, taking into account both the user and the technological point of view.

From the socio-economic perspective, developing components for standardizing the data lakes will enable the use and the reuse of data. Open data, typically published by the Public Sector, is useful for specific information needs and can be easily retrieved, analysed and sometimes merged with other private datasets collected by the Organization. This will enable and improve a better understanding of the scenarios of interest and the capability of data driven decision making.

Risk Level. The risk level of the Action is minimal since: (a) the technology available for managing and analysing data is mature for business applications. Data management and analysis of a large number of heterogeneous data sources from a technical and technological perspective is a possible task, (b) there is a large number of data sources available to experiment using these approaches. A large number of local and international organizations publish their data as open data, thus making possible their use and reuse for the Action purposes and (c) the Action includes researchers from international research teams working in different computer science fields. The research potential expressed by this network is large since complementary expertise and research skills are available.

3. IMPLEMENTATION

3.1. DESCRIPTION OF THE WORK PLAN

3.1.1. DESCRIPTION OF WORKING GROUPS

Three Working Groups will coordinate the research in the Action (WG1-3) and one will manage the activities (WG4). As the simplified Pert Chart in Section 3.1.3 shows, the project is driven by WG1, where users, stakeholders, researchers and practitioners will identify the needs required for implementing the data lake and will turn these needs into research challenges to be addressed by WG2 and WG3. This process will be repeated twice. The first year will be spent in laying solid foundations to the Action by creating a common background and strengthening the relationships among the Action Members. The last year will be spent in consolidating the research and network outcomes. The Gantt Diagram in Section 3.1.2 describes the Action implementation.

[WG1] Designing the data lake.

Task 1.1 User requirements collection
The task will identify and classify the possible users of a data lake (i.e. the Action Stakeholders) and define the high level functionalities and analytic processes required. These users include both applications over the data lake as well as users who administer and maintain the data lake technology and data.

Task 1.2 Collecting datasets and scenarios
The task will collect datasets and scenarios to be used by researchers and practitioners in order to experiment and evaluate the proposed techniques. Scenarios of interest include not only enterprise business data, but also open datasets. Public Administrations will be involved in the definition of scenarios related to the open data they publish.

Task 1.3 Reference Architecture Specification
This task is in charge of specifying a complete functional architecture for data lakes including also guidelines for the privacy-by-design process. The main building blocks for such systems will be defined in terms of the main functionality, expected inputs, outputs, and user interactions.

[WG2] Building the data lake.

Task 2.1 Defining state of the art and challenges.
The aim of the task is to provide a critical analysis of the emergent techniques for profiling data sources, identifying and critically reviewing the most promising approaches and enabling the development of a conceptual map of the open and closed issues related to the research topic.

Task 2.2 Summarizing content and schema of datasets.
The task is in charge of studying the development of automatic, expressive, complete and effective techniques for the creation of summarized descriptions dataset content (and in case schema) for whatever resides in the data lake.

Task 2.3 Linking datasets and descriptions.
The task is in charge of studying the development of automated solutions for data linkage, that will be able to perform across dataset values and also across metadata descriptions, i.e., dataset profiles.

[WG3] Using and managing the data lake.

Task 3.1 Defining state of the art and challenges.
The aim of the task is to provide a critical analysis of the emergent techniques for discovering and exploiting data sources in a lake, to enable autonomous behaviour and self-management of the data lake when appropriate, identifying and critically reviewing the most promising approaches and enabling the development of a conceptual map of the open and closed issues on the research topic.

Task 3.2 Analysing the data lake
The task is in charge of studying the development of advanced user-friendly query methods and analytics functions that will allow data scientists and other users to identify, analyse and exploit the datasets, or the parts of the datasets, that are related to some specific goal they have in mind also by considering the need of privacy management.

Task 3.3 Evaluating the data lake sources
Benchmarks for evaluating and comparing the frameworks available will be studied. Advanced techniques for evaluating the quality of the datasets in the data lake through an analysis of the profiles will be studied in this task.

Task 3.4 Curating the data lake
The task is in charge to develop techniques for curating the data sources in the lake, e.g., exploiting the data lake logs to improve the maintenance and the ability of the system to address the users' needs. This includes also the study of techniques and issues related to the data provenance, governance (including audit and access control) and data protection in a lake.

[WG4] Action Management

Task 4.1 Participating the Action
The task is in charge of (i) defining strategies for improving the active participation of the members; (ii) defining criteria to select eligible participants in WG meetings; (iii) monitoring the participation of the

Members and the involvement of ITCs; (iv) expanding the Action from the initial Network of Proposers, for improving the balance of its components (in terms of stakeholders, gender, and level of experience)

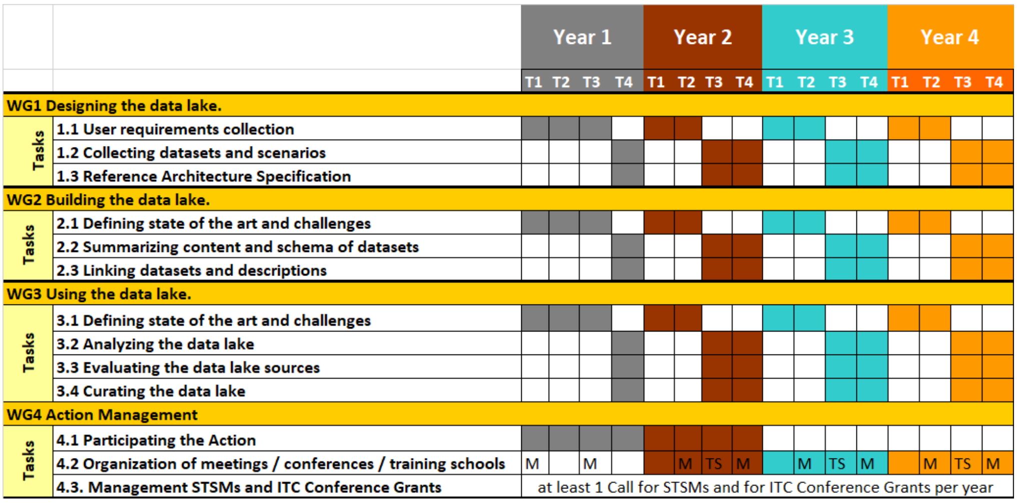
Task 4.2 Organization of meetings / conferences / training schools
The task is in charge of (i) defining topics, agenda and involvement of the Members in the meetings and training schools; (ii) defining the modalities for participating and managing the meeting. Meetings where the attendants will be active will be preferred. Therefore, different activities for stimulating brainstorming, networking and team building will be organized with the understanding that the meetings are the ideal places for creating stable and fruitful relationships.

Task 4.3 Management STSMs and ITC Conference Grants
The task is in charge of defining the criteria to rank the applications for STSMs and ITC Conference Grants, to issue the calls and to propose to the MC the applications to fund.

Action Deliverables:

Based on the activities of the 3 WGs the Action will develop the following deliverables:
D1. Website, with all the Action materials (M3)
D2. Meetings (at least 8 in the Action), including WG Meetings and Open Meetings (e.g., in the form of International conferences).
D3. A flyer describing the Action and its activities, methodology and objectives to be distributed in conferences and other initiatives. An Action Mailing list, a blog, and a twitter communication channel will be created (M6).
D4. Action e-Newsletter (one issue per year)
D5. Training schools with Hackathons (3 in the Action)
D6. A final open conference with international participation (M48)
D7. At least one call per year for STSMs. We plan to fund at least 50 applications. The participant will be asked to write a short scientific report (3 pages) on the activity done and to keep their profile in the website updated with the publications / other outcomes achieved
D8. At least one call per year for ITC Conference Grant
D9. Database of researchers participating in the Action and working in the field
D10. Analysis of the network of researchers (at M12 and updated at M24 and M48).
D11. Organization of at least two workshops in international Conferences (at M48)
D12. Organization of at least two special issues in scientific journals (at M48)
D13. Publications in scientific peer-reviewed journals or conferences authored by Action Members from at least 2 countries (at least 30 at M48).
D14. Bibliography of the literature in the field (M12 and updated during the project).
D15. List of datasets / interesting scenarios (M12 and updated during the project).
D16. Materials for courses and lecturers on the Action topics to be used also for proposing new courses in the existing master programs. (When ready, during the project).
D17. Research raw materials as presentations and reports on the brainstorming sessions organized in the meeting, about the Action research themes (After each meeting).
D18. Final publication including the main research results achieved in the Action and the specification of the reference architecture for a data lake (M48).

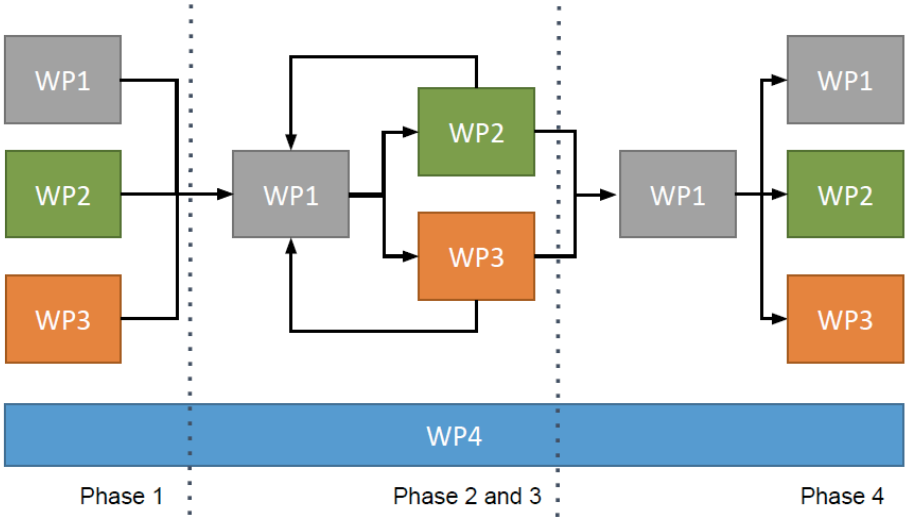
3.1.2. GANTT DIAGRAM



		Year 1				Year 2				Year 3				Year 4			
		T1	T2	T3	T4	T1	T2	T3	T4	T1	T2	T3	T4	T1	T2	T3	T4
WG1 Designing the data lake.																	
Tasks	1.1 User requirements collection	■	■	■		■	■			■	■			■	■		
	1.2 Collecting datasets and scenarios				■			■	■			■	■			■	■
	1.3 Reference Architecture Specification				■			■	■			■	■			■	■
WG2 Building the data lake.																	
Tasks	2.1 Defining state of the art and challenges	■	■	■		■	■			■	■			■	■		
	2.2 Summarizing content and schema of datasets				■			■	■			■	■			■	■
	2.3 Linking datasets and descriptions				■			■	■			■	■			■	■
WG3 Using the data lake.																	
Tasks	3.1 Defining state of the art and challenges	■	■	■		■	■			■	■			■	■		
	3.2 Analyzing the data lake				■			■	■			■	■			■	■
	3.3 Evaluating the data lake sources				■			■	■			■	■			■	■
	3.4 Curating the data lake				■			■	■			■	■			■	■
WG4 Action Management																	
Tasks	4.1 Participating the Action	■	■	■	■	■	■	■	■								
	4.2 Organization of meetings / conferences / training schools	M		M		■	M	TS	M	■	M	TS	M	■	M	TS	M
	4.3. Management STSMs and ITC Conference Grants	at least 1 Call for STSMs and for ITC Conference Grants per year															

Legend: M: WG-MC meetings (we plan to organize joint MC and WG meetings to save money) and Open Meetings; TS: Training school.

3.1.3. PERT CHART (OPTIONAL)



3.1.4. RISK AND CONTINGENCY PLANS

A preliminary evaluation of the Action goals and workplan allowed us to detect only minor risks in the development of working group activities. These risks are related to two main perspectives: 1) research execution / network creation and consolidation, and 2) management of the Action tools.

Description of possible risk	**Impact**	**Prob.**	**Remedial actions**
Research execution / network creation and consolidation			
Lack of data for the definition of the challenges / evaluation of the approaches. Lack of users/ stakeholders for the definition of scenarios	HIGH	LOW	An analysis of the open data available in specific sector will be done. Public Administrations, that typically own data, will be asked to participate in the Action as stakeholders and (open) data providers.
Low quality of the research outcomes / no interest by the research community	LOW	LOW	The proposers include researchers from high quality Institutions who have already a long track record of international publications.
Low diversification of the Action Members in terms of area of expertise, geographical distribution, skills covered ...	HIGH	LOW	The researchers involved already cover a large spectrum of the computer science disciplines. If a lack of expertise in a particular area occurs then experts will be identified and asked to reach the Action as Members or as Invited Speakers in the meetings or Trainers in Training Schools.
Lack of participation from members of industry / data owners, policy and decision makers	HIGH	LOW	Specific calls for inviting people from industry and data owner, policy and decision makers to participate will be issued. Action Members will participate and disseminate the Action activities in events, such as industrial days, organized by their institutions. The connections of the Action Members will be analyzed and relevant possible interested people asked to join the Action.

Lack of active participation by the Action Members in the network activities	HIGH	LOW	The meetings will be conceived as the place where Members have to actively “develop” the research. Diversified strategies for guaranteeing this active participation will be studied and applied.
Management of the Action tools			
Lack of support of MC Members in the definition of the Action strategy.	HIGH	LOW	The ESB will be in charge of planning and executing the Action strategy. The ESB will define strategy for an improvement involvement of the MC Members. In case of low support of come components of the ESB, they will be asked to withdraw and substituted by other members. A direct and open participation in the Action strategy and organization is one of the Action goals and addressed by Task 4.1.
Delay in the production of deliverables	LOW	LOW	Some of the proposers had responsibility roles in other COST Actions. In case of delays, a support from the Action Members will be asked to cover the delay.

3.2. MANAGEMENT STRUCTURES AND PROCEDURES

The Action will follow the rules and procedures of “COST 134/14: COST Action Management, Monitoring and Final Assessment” and will adapt the typical COST Action Management structure:

Management Committee (MC). The MC is in charge of supervising and coordinating the Action. It is composed of not more than two representatives of each participating Country. A **Chair** and a **vice-Chair** of the Action are elected by the MC with the responsibilities defined in the COST Vademecum. Moreover, the MC will appoint:

- a **Scientific Coordinator**, with the responsibility of supporting the MC in the definition and assessment of the research goals to achieve in each Action period;
- a **Dissemination Coordinator**, with the responsibility of monitoring and collecting the dissemination activity promoted by the Action Members;
- a **Working Group Leader** (supported by a co-Leader) for each WG with the role of defining and assuring the achievements of the goals within the established deadline
- a **STSM Coordinator**, with the responsibility of issuing the calls for the STSM application submissions, and ranking them according to the criteria established by the MC;
- a **Training School Coordinator**, in charge of the organization of the Training Schools, by issuing the calls for the selection of the venue, the topics, the Trainers and the trainees to be funded with a grant;
- a **ITC Conference Grant Coordinator**, in charge of the organization, the ranking (according to the criteria established by the MC) and the monitoring of the calls for the selection of members to be granted for participating in conferences.

Executive Scientific Board (ESB). The responsibility for planning and executing of the activities can be delegated by the MC to the ESB, composed of the Chair and the vice-Chair, the Working Group Leaders and the other Coordinators.

Editorial Board. It is composed of a subset of the MC Members and is in charge of planning the publication strategy, reviewing the Action deliverables, and defining the website contents.

There will be one representative person from ITCs for at least one of the responsible roles. The Action encourages and promotes the participation of ECIs not only as researchers and participants to the activities (as WG members, participants in the Training schools, …), but also as Action coordinators (WG Leaders, STSM, Training, Scientific Coordinators,…). This is motivated by two main aspects: first, a large number of researchers involved in the Action proposal are ECIs and, second, the passion and enthusiasm of ECIs may, in case, supply their lack of experience. Nevertheless, it will not constitute a risk for the success of the Action, since Senior researchers will participate in all the activities of the Action and can redress any possible inexperience.

3.3. NETWORK AS A WHOLE

The AMICA Network of Proposers intends to address a hot and challenging topic: studying the technology and software ecosystem of data lake systems for big data management and analysis. Data lakes have been only recently proposed in the literature, but they have created a great deal of interest among academia and industry. This interest is highlighted by the Network of Proposers that includes 130 Members belonging from Organizations located in 28 Countries. The Network has been built around the following principles:

- Including representative members from the three stakeholder categories (researchers; technology providers and vendors; data owners, decision and policy makers). In this way, all the perspectives provided by the stakeholders will be analysed and taken into account both in the specification of the problem requirements and in the definition and evaluation of the possible solutions. Note that Participants in the Network can belong to more than one of these categories. If during the Action it is observed that there is a lack of stakeholders from some specific category then specific remedial actions, as the one described in Section 3.1.4, will be planned and executed.
- Including researchers with expertise in diversified but complementary fields in several computer science disciplines. This is due to the multifaceted nature of the problem, where the data sources are stored in their native form and the management/analysis is postponed until the user expresses a specific information need. For this reason, data lake development requires the involvement of experts with skills able to deal with all the steps of the typical workflows of data management, and analysis. Moreover, the volume and the heterogeneity of the data managed requires jointly applied methodologies, techniques and tools from different areas of computer science. Finally, a systemic approach needs to take into account aspects related to data curation, data governance and the definition of standard components and architectures
- Including researchers from laboratories in different geographical areas and with different professional experiences. This will allow the participants, and in particular ECIs, who have less experience, to share different research approaches and to improve and refine their own research methodology. This will allow also participants from ITCs to benefit from the research infrastructures of other countries, thus improving their ability to promote high quality research.

Members from Network of Proposers have strong relationships with researchers from international qualified research centres located in NNCs or IPCs. These connections will be carefully evaluated and selected experts will be invited to join the Action if there will be a reciprocal benefit from their participation in the Action. For the moment the Network includes Members from one NNC who, thanks to their participation in a previous project, can share a knowledge base composed of massive amounts of structured data and documents for future experimentation.

Finally, the Network of Proposers exhibits the typical (unbalanced) ratio with regards to gender found in computer science (79.2% Males, 20.8% Females), the number of ECIs is 47 out of 130 and have been involved Institutions from 28 countries (50% of them are ITCs). The Institutional distribution of Network of Proposers is 83.8% Higher Education & Associated Organisations, 11.5% Business enterprise, 2.3% Government/Intergovernmental Organisations except Higher Education, 1.5% Private Non-Profit without market revenues, NGO, and 0.8% Standards Organisation.

The Action Members will work for enlarging and improving the balance of the participants. One of the aims of Task 4.1 is to plan and implement strategies for reducing unbalanced ratio in the Action participants.

References

[AGN15] Z. Abedjan, L. Golab, F. Naumann: Profiling relational data: a survey. VLDB J. 24(4): 557-581 2015

[BMNT11] M. Bergman, T. Milo, S. Novgorodov, and W. C. Tan. Query-oriented data cleaning with oracles. In SIGMOD, p. 1199–1214, 2015.

[CFS15] W. Carrara, S. Fischer, E. van Steenbergen: Analytical Report 2: E-skills and Open Data, https://www.europeandataportal.eu/en/highlights/discover-what-skills-are-required-work-open-data

[CM+15] X. Chu, J. Morcos, I. F. Ilyas, M. Ouzzani, P. Papotti, N. Tang, and Y. Ye. KATARA: A data cleaning system powered by knowledge bases and crowdsourcing. In SIGMOD, p. 1247–1261, 2015.

[DFA+17] D. Deng, R. C. Fernandez, Z. Abedjan, et al: The Data Civilizer System. CIDR 2017
[EB11] S. Elbassuoni, R. Blanco. Keyword search over RDF graphs. In CIKM 2011

[EIV07] Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, and Vassilios S. Verykios. Duplicate Record Detection: A Survey. IEEE Transactions on Knowledge and Data Engineering, 19(1):1–16, 2007.

[F17] Andrew Cave What Will We Do When The World's Data Hits 163 Zettabytes In 2025?, Forbes 2017 https://www.forbes.com/sites/andrewcave/2017/04/13/what-will-we-do-when-the-worlds-data-hits-163-zettabytes-in-2025

[FAM+16] R. C. Fernandez, Z. Abedjan, S. Madden, M. Stonebraker: Towards large-scale data discovery: position paper. ExploreDB@SIGMOD/PODS 2016: 3-5

[FGW11] W. Fan, F. Geerts, and J. Wijsen. Determining the currency of data. In PODS, p. 71–82, 2011.

[FLW11] J. Feng, G. Li, and J. Wang, “Finding top-k answers in keyword search over relational databases using tuple units,” IEEE Trans. Knowl. Data Eng., vol. 23, no. 12, pp. 1781–1794, 2011.

[FRI+15]M. Farid, A. Roatis, Ihab F. Ilyas, Hella-Franziska Hoffmann, and Xu Chu. 2016. CLAMS: Bringing Quality to Data Lakes. SIGMOD 2016.

[GMQ17] Gartner 2017 Magic Quadrant for Data Science Platforms
https://www.kdnuggets.com/2017/02/gartner-2017-mq-data-science-platforms-gainers-losers.html

[HGQ16] R. Hai, S. Geisler, and C. Quix. Constance: An intelligent data lake system. In SIGMOD, pages 2097–2100, 2016.

[J09] T.Johnson: Data Profiling. Encyclopedia of Database Systems 2009: 604-608

[KHW+17] S. Kruse, D. Hahn, M. Walter, F. Naumann: Metacrate: Organize and Analyze Millions of Data Profiles. CIKM 2017: 2483-2486

[KKA+17] N. Konstantinou, Martin Koehler, E. Abel, et al. The VADA Architecture for Cost-Effective Data Wrangling. SIGMOD 2017.

[LBL16] H. Lakkaraju, S. H. Bach, J. Leskovec: Interpretable Decision Sets: A Joint Framework for Description and Prediction. KDD 2016: 1675-1684

[LS16] A. LaPlante, B. Sharma: Architecting Data Lakes, O'Reilly 2016

[MLPV16] D. Mottin, M. Lissandrini, T. Palpanas and Y. Velegrakis, "Exemplar Queries: A New Way of Searching", VLDBJ, 25(6), 2016.

[MT17] A. Maccioni, R. Torlone: Crossing the finish line faster when paddling the Data Lake with Kayak. PVLDB 10(12): 1853-1856, 2017

[ODM16] Open Data Maturity in Europe, European data portal Report, 2016. https://www.europeandataportal.eu/sites/default/files/edp_landscaping_insight_report_n2_2016.pdf

[SP16] I. Suriarachchi, B. Plale: Provenance as Essential Infrastructure for Data Lakes. IPAW 2016: 178-182

[SV07] D. Srivastava, Y. Velegrakis:Intensional associations between data and metadata. SIGMOD 2007: 401-412

[WM13] E. Wu, S. Madden: Scorpion: Explaining Away Outliers in Aggregate Queries. PVLDB 6(8): 553-564 (2013)

[ZM16] C. Zhai, S. Massung: Text Data Management and Analysis: a Practical Introduction to Information Retrieval and Text Mining. ACM and Morgan & Claypool, USA, 2016.

[YQC10] J. X. Yu, L. Qin, and L. Chang, Keyword Search in Databases. Morgan & Claypool, 2010

COST Mission and Policies

The AMICA Action is conceived to fully address the COST Mission of enabling breakthrough scientific developments. This is achieved thanks to two main ingredients: a challenging and hot research topic with a potential impact in several real scenarios and a network of researchers with complementary expertise.

The development and use of data lakes for the management and analysis of big data is now at the early stage. Some preliminary research projects have been proposed, and some commercial tools have been developed. Nevertheless, a common vision, a reference architecture, an agreed set of expected components and functionalities, a number of benchmarks and criteria for the performance evaluation are still missing. The implementation of a COST Action on this topic is timely: the research effort and the scientific / coordination and networking outcomes can now play a central and decisive in the (European) research landscape.

The network of proposers includes: i) researchers in complementary computer science disciplines. This will make the Action able to deal with different data source formats, to apply techniques from different computer science areas, to address the problem with different methodologies; ii) data owners. They aim to fully exploit the potential of their data assets and can provide scenarios, needs and use cases; iii) companies. They are interested in the exploitation of the results achieved and will provide the “business” perspective on the problem. To be more focused on the technological and technical issues, the network does not deliberately include in this initial phase policy and decision makers nor researchers from the economics disciplines. Nevertheless, we are aware that their contribution is needed and representative people from these fields will be invited at least as experts in the meetings.

To achieve its Mission, the AMICA Action will fully exploit all the networking tools provided by the COST Framework, addressing and supporting the COST policies and rules as follows:

1. **Promoting an “open and inclusive participation” in all the Action networking activities**. COST promotes the development of open Networks. AMICA, not only will be open, but also will allow a large number of participants to assume an active role by: a) implementing a mechanism based on “calls for application” for enabling a large involvement of the WG Members in the organization and management of the activities (e.g., through calls for meeting organization and / or hosting; calls for the organization of the training schools, calls for participation in the events). The calls will ask for applications where the organization committee and the main responsible people include ECIs and researchers from ITCs. b) doubling the WG leadership roles (with a leader and a co-leader, as reported in the technical annex) we will strive to involve more ECIs and ITCs in the responsibility positions. c) preferring open large meetings to small meetings. In this way, the participants can have more opportunities to networking and build new research relationships. The meetings will include brainstorming sessions / team building activities / working sessions to make participants active and working together, thus enabling the creation of new projects and new research opportunities. d) defining criteria for distributing and balancing the grants and the benefits among the participants and the countries, thus promoting the participation of ECI’s, people from ITCs and gender balance.
2. **Supporting the dissemination of results of such research activities**. The dissemination of the outcomes is a critical activity, since it contributes to the enlargement of the Action to other interested researchers and in making the Action as a reference point in the research community. The participants will be encouraged from the beginning to starting research activity together and acknowledging the Action in their dissemination. AMICA will directly organize open meetings (in the form of International Conferences) where it will be possible to disseminate part of the achieved research outcomes. Moreover, Action Members intend to promote special issues in international journals, and workshops in International Conferences that will promote AMICA out of the Action Members.
3. **Supporting ITC researchers to join the Action.** Besides the activities fostering the participation from ITCs described in point 1, AMICA will address this policy by trying to balance the number of researchers from ITCs and non-ITCs who will benefit from the Action networking tools (through targeted invitations, involving the MC Members, reserving grants, …). Moreover, if and when possible, the meetings, and the training schools will be organized in ITCs, thus promoting the participation of local researchers and the visibility of the Action (and of the COST framework). The MC will give higher priority to STSM applications involving

ITCs, and call for assigning conference grants to ITC researchers will be issued.

4. **Supporting the participation of Non-COST Countries in COST activities**. The network of proposers already includes one non-COST Country which owns a large repository of data sources. The research connections of Proposers in the Network will be furthermore analyzed and the institutes from NNCs where we found a reciprocal benefit generated from the participation will be invited to participate in the Action.

Network of Proposers - Features

COST Inclusiveness target countries

50.00 %

Number of Proposers

130

Geographic Distribution of Proposers

Country	ITC/ non ITC/ other	Number of institutions from that country	Number of researchers from that country	Percentage of the proposing network
Austria	non ITC	2	2	1.54 %
Belgium	non ITC	2	2	1.54 %
Bulgaria	ITC	1	1	0.77 %
Croatia	ITC	11	11	8.46 %
Cyprus	ITC	2	3	2.31 %
Czech Republic	ITC	1	1	0.77 %
Estonia	ITC	1	1	0.77 %
Finland	non ITC	1	1	0.77 %
France	non ITC	12	12	9.23 %
Germany	non ITC	5	5	3.85 %
Greece	non ITC	3	3	2.31 %
Ireland	non ITC	3	3	2.31 %
Israel	non ITC	1	1	0.77 %
Italy	non ITC	13	14	10.77 %
Japan	other	1	1	0.77 %
Malta	ITC	3	3	2.31 %
Netherlands	non ITC	7	7	5.38 %
Norway	non ITC	1	1	0.77 %
Poland	ITC	6	6	4.62 %
Portugal	ITC	2	2	1.54 %
Romania	ITC	5	5	3.85 %
Serbia	ITC	4	4	3.08 %
Slovakia	ITC	5	5	3.85 %
Slovenia	ITC	4	4	3.08 %
Spain	non ITC	17	17	13.08 %
Switzerland	non ITC	3	3	2.31 %
Turkey	ITC	2	2	1.54 %
United Kingdom	non ITC	7	8	6.15 %
fYR Macedonia	ITC	2	2	1.54 %

Gender Distribution of Proposers

79.2% Males
20.8% Females

Average Number of years elapsed since PhD graduation of Proposers with a doctoral degree

12.8

Number of Early Career Investigators

47

Core Expertise of Proposers: Distribution by Sub-Field of Science

63.8% Computer and Information Sciences
25.4% Electrical engineering, electronic engineering, Information engineering
1.5% Biological sciences
0.8% Earth and related Environmental sciences
0.8% Economics and business
4% Other
3.8% Unspecified

Institutional distribution of Network of Proposers

83.8% Higher Education & Associated Organisations
11.5% Business enterprise
2.3% Government/Intergovernmental Organisations except Higher Education
1.5% Private Non-Profit without market revenues, NGO
0.8% Standards Organisation

Higher Education & Associated Organisations:109

- Number by Field of Science of Department/Faculty of Affiliation
  Electrical engineering, electronic engineering, Information engineering:24
  Computer and Information Sciences:77
  Other social sciences:1
  Earth and related Environmental sciences:1
  Languages and literature:1
  Other humanities:1
  Interdisciplinary:1
  Mathematics:1
- Number by Type
  Research Oriented:49
  Education Oriented:60
- Number by Ownership
  Fully or mostly public:103
  50-50 Public and Private:2
  Fully or mostly private:4

Private Non-Profit without market revenues, NGO:2

- Number by Type
  Other:1
  Charity:1
- Number by Level
  National:2

Government/Intergovernmental Organisations except Higher Education:3

- Number by Level

Central and Federal Government:2
Local government:1

- Number by Type
  R&D Funding and/or R&D Performing bodies:2
  Other Public Non-Profit Institution:1

Business enterprise:15

- Number by Market sector of unit of affiliation
  Information And Communication:11
  Professional, Scientific And Technical Activities:4
- Number by Type
  Private enterprises:14
  Research and Technology Organization - RTO:1
- Number by Ownership and International Status
  Enterprise owned by a foreign multinational group:2
  Independent Enterprise:13
- Number by Size
  Large company:2
  SME (EU Definition provided underneath after selection):13

Standards Organisation:1

- Number by Membership type
  With no government membership:1
- Number by Level
  International:1

COST Country Institutions(28) : Austria , Belgium , Bulgaria , Croatia , Cyprus , Czech Republic , Estonia , Finland , France , Germany , Greece , Ireland , Israel , Italy , Malta , Netherlands , Norway , Poland , Portugal , Romania , Serbia , Slovakia , Slovenia , Spain , Switzerland , Turkey , United Kingdom , fYR Macedonia
Near-Neighbour Country Institutions(0)
COST International Partners(2) : China, Japan
European Commission and EU Agencies(0)
European RTD Organisations(0)
International Organisations(0)

Network of Proposers - Details

Main Proposer's Details

Title:	Prof	**Gender:**	M
First Name:	Francesco	**Year of birth:**	21/05/1973
Last Name:	Guerra	**Years from PhD:**	14
Email:	francesco.guerra@unimore.it	**Telephone Number:**	+390592056264
Institution:	Università di Modena e Reggio Emilia	**Type of Institution:**	Higher Education & Associated Organisations
Address of the Institution:	Via Vivarelli 10, 41125 Modena, Italy		
Sub-field of Science of Department:	Electrical engineering, electronic engineering, Information engineering	**Core Area of Expertise:**	Computer and Information Sciences (Theoretical aspects of data curation, data mining and database handling)

Secondary Proposers' Details

Austria

Dr Mihai Lupu (Research Studios Austria [Studio Data Science])

Participating as Secondary Proposer
E-mail: lupu@ifs.tuwien.ac.at
Telephone: +436804012669
Core Expertise: Computer and Information Sciences: Information Retrieval
Gender: M
Years from PhD: 10

Dr Javier D Fernandez (Vienna University of Economics and Business)

Participating as Secondary Proposer
E-mail: javier.fernandez@wu.ac.at
Telephone: +43067761372217
Core Expertise: Computer and Information Sciences: Big (Semantic) Data Management
Gender: M
Years from PhD: 4

Belgium

Prof Stijn Vansummeren (Université Libre de Bruxelles)

Participating as Secondary Proposer
E-mail: svsummer@ulb.ac.be
Telephone: +32497946534
Core Expertise: Computer and Information Sciences: Theoretical aspects of data curation, data mining and database handling
Gender: M
Years from PhD: 13

Prof Jef Wijsen (UMONS [Département d'Informatique])

Participating as Secondary Proposer
E-mail: jef.wijsen@umons.ac.be
Telephone: +32472428155
Core Expertise: Computer and Information Sciences: Database systems
Gender: M
Years from PhD: 23

Bulgaria

Prof Velislava Stoykova (Institute for Bulgarian Language)

Participating as Secondary Proposer
E-mail: vstoykova@yahoo.com
Telephone: +35929792942
Core Expertise: Languages and literature: Databases, data mining, data curation, computational modelling
Gender: F
Years from PhD: 14

Croatia

Prof Sanda Martincic Ipsic (University of Rijeka [Department of informatics])

Participating as Secondary Proposer
E-mail: smarti@inf.uniri.hr
Telephone: +38551584714
Core Expertise: Computer and Information Sciences: Artificial intelligence, intelligent systems, multi agent systems
Gender: F
Years from PhD: 11

Prof Markus Schatten (University of Zagreb - Fakultet organizacije i informatike)
Participating as Secondary Proposer
E-mail: markus.schatten@foi.hr
Telephone: +385981648617
Core Expertise: Computer and Information Sciences: Artificial intelligence, intelligent systems, multi agent systems
Gender: M
Years from PhD: 8

Prof Kornelije Rabuzin (University of Zagreb - Faculty of organization and informatics [Department for theoretical and applied foundations of information science])
Participating as Secondary Proposer
E-mail: kornelije.rabuzin@foi.hr
Telephone: 0038542390836
Core Expertise: Computer and Information Sciences: Theoretical aspects of data curation, data mining and database handling
Gender: M
Years from PhD: 11

Prof Sandra Lovrenčić (University of Zagreb - Faculty of organization and informatics [Department of theoretical and applied foundations of information sciences])
Participating as Secondary Proposer
E-mail: sandra.lovrencic@foi.hr
Telephone: +38542390851
Core Expertise: Computer and Information Sciences: Artificial intelligence, intelligent systems, multi agent systems
Gender: F
Years from PhD: 11

Prof Ana Mestrovic (University of Rijeka)
Participating as Secondary Proposer
E-mail: amestrovic@inf.uniri.hr
Telephone: +38551584716
Core Expertise: Computer and Information Sciences: Artificial intelligence, intelligent systems, multi agent systems
Gender: F
Years from PhD: 9

Dr Marina Ivasic-Kos (Department of Informatics University of Rijeka)
Participating as Secondary Proposer
E-mail: marinai@uniri.hr
Telephone: +38551584700
Core Expertise: Electrical engineering, electronic engineering, Information engineering: Computer vision
Gender: F
Years from PhD: 6

Dr Marko Horvat (Zagreb University of Applied Sciences [Computer Science and Information Technology Department])
Participating as Secondary Proposer
E-mail: marko.horvat@tvz.hr
Telephone: +38515603944
Core Expertise: Computer and Information Sciences: Artificial intelligence, intelligent systems, multi agent systems
Gender: M
Years from PhD: 5

Mr Sergej Lugovic (Zagreb University of Applied Sciences)

Participating as Secondary Proposer
E-mail: lugovicsergej@gmail.com
Telephone: 00385914658199
Core Expertise: Computer and Information Sciences: Artificial intelligence, intelligent systems, multi agent systems
Gender: M
Years from PhD: 0

Dr Jasminka Dobsa (University of Zagreb - Faculty of Organization and Informatics [Laboratory for Generative Programming and Machine Learning])
Participating as Secondary Proposer
E-mail: jasminka.dobsa@foi.hr
Telephone: +385 42 3390844
Core Expertise: Computer and Information Sciences: Mathematics applied to computer science, mathematical aspects of computer science
Gender: F
Years from PhD: 12

Mr Marko Štajcer (Poslovna inteligencija d.o.o. [Innovation & Development])
Participating as Secondary Proposer
E-mail: marko.stajcer@inteligencija.com
Telephone: 00385992202962
Core Expertise: Electrical engineering, electronic engineering, Information engineering: Databases, data mining, data curation, computational modelling
Gender: M
Years from PhD: 0

Mr Davorin Cetto (Syntio d.o.o.)
Participating as Secondary Proposer
E-mail: davorin.cetto@syntio.hr
Telephone: +385913212980
Core Expertise: Computer and Information Sciences: Machine learning algorithms
Gender: M
Years from PhD: 0

Cyprus

Dr Georgia Kapitsaki (University of Cyprus)
Participating as Secondary Proposer
E-mail: gkapi@cs.ucy.ac.cy
Telephone: 0035722892692
Core Expertise: Computer and Information Sciences: Theoretical aspects of pervasive and ubiquitous computing
Gender: F
Years from PhD: 9

Dr Loizos Michael (Open University of Cyprus)
Participating as Secondary Proposer
E-mail: loizos@ouc.ac.cy
Telephone: +35722411963
Core Expertise: Computer and Information Sciences: Artificial intelligence, intelligent systems, multi agent systems
Gender: M
Years from PhD: 10

Dr Ekaterini Ioannou (Open University of Cyprus)
Participating as Secondary Proposer
E-mail: katerina.ioannou@gmail.com
Telephone: +35722411966

Core Expertise: Computer and Information Sciences: Theoretical aspects of data curation, data mining and database handling
Gender: F
Years from PhD: 7

Czech Republic

Dr Filip Zavoral (Charles University - http://www.ksi.mff.cuni.cz/~zavoral/ [Faculty of Mathematics and Physics, Department of Software Engineering])

Participating as Secondary Proposer
E-mail: zavoral@ksi.mff.cuni.cz
Telephone: 951544250
Core Expertise: Computer and Information Sciences: Algorithms, distributed, parallel and network algorithms
Gender: M
Years from PhD: 21

Estonia

Dr Innar Liiv (Tallinn University of Technology)

Participating as Secondary Proposer
E-mail: innar.liiv@ttu.ee
Telephone: +3725200552
Core Expertise: Computer and Information Sciences: Artificial intelligence, intelligent systems, multi agent systems
Gender: M
Years from PhD: 10

Finland

Prof Vagan Terziyan (University of Jyvaskyla)

Participating as Secondary Proposer
E-mail: vagan.terziyan@jyu.fi
Telephone: +358503732127
Core Expertise: Computer and Information Sciences: Artificial intelligence, intelligent systems, multi agent systems
Gender: M
Years from PhD: 25

France

Prof Omar BOUCELMA (AIX MARSEILLE UNIVERSITE)

Participating as Secondary Proposer
E-mail: omar.boucelma@univ-amu.fr
Telephone: +33491056053
Core Expertise: Computer and Information Sciences: Theoretical aspects of data curation, data mining and database handling
Gender: M
Years from PhD: 22

Dr Genoveva Vargas-Solar (Centre National de la Recherche Scientifique, Laboratoire LIG)

Participating as Secondary Proposer
E-mail: genoveva.vargas@imag.fr
Telephone: +33476827248
Core Expertise: Computer and Information Sciences: Theoretical aspects of data curation, data mining and database handling
Gender: F
Years from PhD: 13

Prof Catarina Ferreira da Silva (Université Claude Bernard Lyon 1)

Participating as Secondary Proposer

E-mail: catarina.ferreira@univ-lyon1.fr
Telephone: +33426234446
Core Expertise: Computer and Information Sciences: Service Science and Cloud Computing
Gender: F
Years from PhD: 11

Prof themis palpanas (paris descartes university)
Participating as Secondary Proposer
E-mail: themis@mi.parisdescartes.fr
Telephone: +33183945865
Core Expertise: Computer and Information Sciences: Algorithms, distributed, parallel and network algorithms
Gender: M
Years from PhD: 15

Dr Khalid Belhajjame (University Paris-Dauphine)
Participating as Secondary Proposer
E-mail: kbelhajj@googlemail.com
Telephone: +33695084118
Core Expertise: Computer and Information Sciences: Information and Knowledge Management
Gender: M
Years from PhD: 14

Prof Bogdan CAUTIS (University of Paris Sud)
Participating as Secondary Proposer
E-mail: bogdan.cautis@u-psud.fr
Telephone: +33676106285
Core Expertise: Computer and Information Sciences: Algorithms, distributed, parallel and network algorithms
Gender: M
Years from PhD: 11

Ms Ioana MANOLESCU (Inria)
Participating as Secondary Proposer
E-mail: ioana.manolescu@inria.fr
Telephone: +33172925920
Core Expertise: Computer and Information Sciences: Theoretical aspects of data curation, data mining and database handling
Gender: F
Years from PhD: 17

Dr Mahmoud Barhamgi (Claude Bernard Lyon 1 University)
Participating as Secondary Proposer
E-mail: mahmoud.barhamgi@univ-lyon1.fr
Telephone: +3395777734
Core Expertise: Electrical engineering, electronic engineering, Information engineering: Pervasive and ubiquitous computing
Gender: M
Years from PhD: 8

Prof Jérôme Darmont (Université Lumière Lyon 2 [Laboratoire ERIC (EA 3083)])
Participating as Secondary Proposer
E-mail: jerome.darmont@univ-lyon2.fr
Telephone: +33 478 774 403
Core Expertise: Computer and Information Sciences: Databases, Data warehouses
Gender: M
Years from PhD: 19

Prof Antoine Doucet (University of La Rochelle)
Participating as Secondary Proposer
E-mail: antoine.doucet@univ-lr.fr
Telephone: +330546513973
Core Expertise: Computer and Information Sciences: Theoretical aspects of data curation, data mining and database handling
Gender: M
Years from PhD: 13

Dr Fatma ABDELHEDI (TRIMANE [CBI²])
Participating as Secondary Proposer
E-mail: Fatma.Abdelhedi@trimane.fr
Telephone: 0172550068
Core Expertise: Electrical engineering, electronic engineering, Information engineering: Databases, data mining, data curation, computational modelling
Gender: F
Years from PhD: 4

Dr Thomas GERBAUD (OceanData)
Participating as Secondary Proposer
E-mail: contact@oceandata.io
Telephone: 06 48 43 96 56
Core Expertise: Computer and Information Sciences: Theory of scientific computing and data processing
Gender: M
Years from PhD: 13

Germany

Prof Andreas Nürnberger (Otto-von-Guericke University Magdeburg [Faculty of Computer Science])
Participating as Secondary Proposer
E-mail: andreas.nuernberger@ovgu.de
Telephone: +493916758487
Core Expertise: Computer and Information Sciences: Unspecified
Gender: M
Years from PhD: 17

Dr Elena Demidova (Leibniz Universität Hannover)
Participating as Secondary Proposer
E-mail: demidova@L3S.de
Telephone: +49 511 762 17776
Core Expertise: Computer and Information Sciences: Artificial intelligence, intelligent systems, multi agent systems
Gender: F
Years from PhD: 5

Dr Stefan Dietze (L3S Research Center, Leibniz University Hannover)
Participating as Secondary Proposer
E-mail: dietze@l3s.de
Telephone: +491795939815
Core Expertise: Computer and Information Sciences: Artificial intelligence, intelligent systems, multi agent systems
Gender: M
Years from PhD: 14

Dr Davide Mottin (Hasso Plattner Institute [Knowledge Discovery and Data Mining])
Participating as Secondary Proposer
E-mail: davide.mottin@hpi.de

Telephone: +4933155091374
Core Expertise: Computer and Information Sciences: Theoretical aspects of data curation, data mining and database handling
Gender: M
Years from PhD: 3

Prof Jorge Cardoso (Huawei Research - Huawei Munich Research Center [IT Software Infrastructure Laboratory GRC])

Participating as Secondary Proposer
E-mail: jorge.cardoso@huawei.com
Telephone: +49 172 174 0862
Core Expertise: Computer and Information Sciences: Algorithms, distributed, parallel and network algorithms
Gender: M
Years from PhD: 16

Greece

Prof Manolis Wallace (University of Peloponnese)

Participating as Secondary Proposer
E-mail: wallace@uop.gr
Telephone: +306974497183
Core Expertise: Electrical engineering, electronic engineering, Information engineering: Knowledge engineerng
Gender: M
Years from PhD: 13

Dr Foivos Mylonas (Ionian University - Ionian University [Department of Informatics])

Participating as Secondary Proposer
E-mail: fmylonas@ionio.gr
Telephone: 00302661087741
Core Expertise: Computer and Information Sciences: Theoretical aspects of data curation, data mining and database handling
Gender: M
Years from PhD: 10

Dr Dimitris Skoutas (Research Center "Athena" - Institute for the Management of Information Systems)

Participating as Secondary Proposer
E-mail: dskoutas@imis.athena-innovation.gr
Telephone: +302106875403
Core Expertise: Electrical engineering, electronic engineering, Information engineering: Databases, data mining, data curation, computational modelling
Gender: M
Years from PhD: 10

Ireland

Dr Jeremy Debattista (Trinity College Dublin)

Participating as Secondary Proposer
E-mail: debattij@tcd.ie
Telephone: +353894141761
Core Expertise: Computer and Information Sciences: Linked Data and Semantic Web
Gender: M
Years from PhD: 1

Dr John Breslin (National University of Ireland Galway)

Participating as Secondary Proposer
E-mail: john.breslin@nuigalway.ie

Telephone: +35391492622
Core Expertise: Electrical engineering, electronic engineering, Information engineering: Semantic Web and Social Media
Gender: M
Years from PhD: 16

Dr Abdulhussain Mahdi (University of Limerick)

Participating as Secondary Proposer
E-mail: hussain.mahdi@ul.ie
Telephone: +35361213492
Core Expertise: Electrical engineering, electronic engineering, Information engineering: Development of scientific computing, data processing, simulation and modelling tools
Gender: M
Years from PhD: 28

Israel

Dr Chaya Liebeskind (Jerusalem College of Technology)

Participating as Secondary Proposer
E-mail: liebchaya@gmail.com
Telephone: +972777848477
Core Expertise: Computer and Information Sciences: Natural Language Processing
Gender: F
Years from PhD: 2

Italy

Prof Yannis Velegrakis (University of Trento)

Participating as Secondary Proposer
E-mail: velgias@disi.unitn.eu
Telephone: +390461283986
Core Expertise: Computer and Information Sciences: Theory of scientific computing and data processing
Gender: M
Years from PhD: 13

Mr Matteo Paganelli (Università degli studi di Modena e Reggio Emilia [Ingegneria])

Participating as Secondary Proposer
E-mail: matteo.paganelli@unimore.it
Telephone: +059335388
Core Expertise: Electrical engineering, electronic engineering, Information engineering: Databases, data mining, data curation, computational modelling
Gender: M
Years from PhD: 0

Dr Laura Po (Università di Modena e Reggio Emilia)

Participating as Secondary Proposer
E-mail: laura.po@unimore.it
Telephone: +390592056255
Core Expertise: Electrical engineering, electronic engineering, Information engineering: Databases, data mining, data curation, computational modelling
Gender: F
Years from PhD: 9

Prof Federica Mandreoli (Università di Modena e Reggio Emilia)

Participating as Secondary Proposer
E-mail: federica.mandreoli@unimo.it
Telephone: +390592056134
Core Expertise: Computer and Information Sciences: Database and Information Systems

Gender: F
Years from PhD: 17

Dr Marco Brambilla (Politecnico di Milano)
Participating as Secondary Proposer
E-mail: marco.brambilla@polimi.it
Telephone: +390223997621
Core Expertise: Electrical engineering, electronic engineering, Information engineering: Databases, data mining, data curation, computational modelling
Gender: M
Years from PhD: 13

Prof Riccardo Torlone (Universita' Roma Tre)
Participating as Secondary Proposer
E-mail: torlone@dia.uniroma3.it
Telephone: +390657333377
Core Expertise: Electrical engineering, electronic engineering, Information engineering: Databases, data mining, data curation, computational modelling
Gender: M
Years from PhD: 28

Prof Nicola Ferro (University of Padua)
Participating as Secondary Proposer
E-mail: ferro@dei.unipd.it
Telephone: +390498277939
Core Expertise: Computer and Information Sciences: Information Retrieval
Gender: M
Years from PhD: 13

Dr Mauro Dragoni (Fondazione Bruno Kessler)
Participating as Secondary Proposer
E-mail: dragoni@fbk.eu
Telephone: +393486861606
Core Expertise:
Gender: M
Years from PhD: 8

Dr Davide Taibi (National Research Council of Italy)
Participating as Secondary Proposer
E-mail: davide.taibi@itd.cnr.it
Telephone: +390916809216
Core Expertise: Educational sciences: Databases, data mining, data curation, computational modelling
Gender: M
Years from PhD: 0

Prof Andrea Maurino (Università degli studi di Milano - Bicocca [dipartimento di informatica, sistemistica e comunicazione])
Participating as Secondary Proposer
E-mail: maurino@disco.unimib.it
Telephone: +390264487897
Core Expertise: Electrical engineering, electronic engineering, Information engineering: Databases, data mining, data curation, computational modelling
Gender: M
Years from PhD: 16

Dr Mirjana Mazuran (IIT - Human Technopole)
Participating as Secondary Proposer

E-mail: mirjana.mazuran@polimi.it
Telephone: +393487494180
Core Expertise: Electrical engineering, electronic engineering, Information engineering: Databases, data mining, data curation, computational modelling
Gender: F
Years from PhD: 0

Prof Fabrizio Riguzzi (Università di Ferrara - Dipartimento di Matematica e Informatica)
Participating as Secondary Proposer
E-mail: fabrizio.riguzzi@unife.it
Telephone: +39 0532974792
Core Expertise: Computer and Information Sciences: Machine learning algorithms
Gender: M
Years from PhD: 19

Mr Claudio Montanari (KNOW-HOW SrL [Management])
Participating as Secondary Proposer
E-mail: claudio.montanari@know-how.it
Telephone: +393487307217
Core Expertise: Electrical engineering, electronic engineering, Information engineering: Databases, data mining, data curation, computational modelling
Gender: M
Years from PhD: 22

Japan

Mr David Sanz (everis [Data & Analytics])
Participating as Secondary Proposer
E-mail: david.sanz.martinez@everis.com
Telephone: +34 917 49 00 00
Core Expertise: Computer and Information Sciences: Theoretical aspects of data curation, data mining and database handling
Gender: M
Years from PhD: 0

Malta

Dr Colin Layfield (University of Malta)
Participating as Secondary Proposer
E-mail: colin.layfield@um.edu.mt
Telephone: +35623402515
Core Expertise: Computer and Information Sciences: Artificial intelligence, intelligent systems, multi agent systems
Gender: M
Years from PhD: 15

Dr Charlie Abela (University of Malta - Department of Artificial Intelligence [Faculty of ICT])
Participating as Secondary Proposer
E-mail: charlie.abela@um.edu.mt
Telephone: +35623402027
Core Expertise: Computer and Information Sciences: Artificial intelligence, intelligent systems, multi agent systems
Gender: M
Years from PhD: 2

Dr Joel Azzopardi (University of Malta [Department of Intelligent Computer Systems])
Participating as Secondary Proposer
E-mail: joel.azzopardi@um.edu.mt
Telephone: +35623403541

Core Expertise: Computer and Information Sciences: Artificial intelligence, intelligent systems, multi agent systems
Gender: M
Years from PhD: 6

Netherlands

Dr Odysseas Papapetrou (Eindhoven University of Technology [Databases group])
Participating as Secondary Proposer
E-mail: odysseasp@gmail.com
Telephone: +357 99 906644
Core Expertise: Computer and Information Sciences: Algorithms, distributed, parallel and network algorithms
Gender: M
Years from PhD: 7

Dr George Fletcher (Eindhoven University of Technology)
Participating as Secondary Proposer
E-mail: g.h.l.fletcher@tue.nl
Telephone: +31402472624
Core Expertise: Computer and Information Sciences: Theory of scientific computing and data processing
Gender: M
Years from PhD: 11

Dr Dimitrios Koureas (Naturalis Biodiversity Center [International Cooperation on Biodiversity Infrastructures])
Participating as Secondary Proposer
E-mail: dimitris.koureas@naturalis.nl
Telephone: +31717519251
Core Expertise: Electrical engineering, electronic engineering, Information engineering: Biodiversity Informatics
Gender: M
Years from PhD: 6

Mr Wouter Addink (Naturalis Biodiversity Center)
Participating as Secondary Proposer
E-mail: wouter.addink@naturalis.nl
Telephone: +31717517364
Core Expertise: Biological sciences: Bioinformatics
Gender: M
Years from PhD: 0

Dr Peter Schalk (Species 2000)
Participating as Secondary Proposer
E-mail: peter.schalk@naturalis.nl
Telephone: +31717517362
Core Expertise: Biological sciences: Environmental and marine biology
Gender: M
Years from PhD: 30

Dr Asterios Katsifodimos (Delft University of Technology [Web Information Systems Group])
Participating as Secondary Proposer
E-mail: a.katsifodimos@tudelft.nl
Telephone: +31 6 29 35 10 89
Core Expertise: Computer and Information Sciences: Theoretical aspects of data curation, data mining and database handling
Gender: M
Years from PhD: 5

Prof Geert-Jan Houben (TU Delft)

Participating as Secondary Proposer
E-mail: g.j.p.m.houben@tudelft.nl
Telephone: +31152787486
Core Expertise: Computer and Information Sciences: Web engineering, databases, user modeling
Gender: M
Years from PhD: 28

Norway

Prof Kjetil Norvag (Norwegian University of Science and Technology)

Participating as Secondary Proposer
E-mail: Kjetil.Norvag@idi.ntnu.no
Telephone: +4773596755
Core Expertise: Computer and Information Sciences: Theoretical aspects of data curation, data mining and database handling
Gender: M
Years from PhD: 18

Poland

Dr Julian Szymanski (Gdansk University of Technology)

Participating as Secondary Proposer
E-mail: julian.szymanski@eti.pg.gda.pl
Telephone: +48501942784
Core Expertise: Computer and Information Sciences: Artificial intelligence, intelligent systems, multi agent systems
Gender: M
Years from PhD: 9

Prof Ngoc Thanh Nguyen (Wroclaw University of Science and Technology [Department of Information Systems (W8/K1)])

Participating as Secondary Proposer
E-mail: Ngoc-Thanh.Nguyen@pwr.edu.pl
Telephone: +48713204139
Core Expertise: Electrical engineering, electronic engineering, Information engineering: Databases, data mining, data curation, computational modelling
Gender: M
Years from PhD: 16

Prof Grzegorz J. Nalepa (AGH University of Science and Technology [Applied Computer Science])

Participating as Secondary Proposer
E-mail: grzegorz.j.nalepa@uj.edu.pl
Telephone: +48603952846
Core Expertise: Computer and Information Sciences: Artificial intelligence, intelligent systems, multi agent systems
Gender: M
Years from PhD: 14

Dr Szymon Bobek (AGH University of Science and Technology - Faculty of Electrical Engineering, Automatics, Computer Science and Biomedical Engineering)

Participating as Secondary Proposer
E-mail: szymon.bobek@agh.edu.pl
Telephone: +48126173941
Core Expertise: Computer and Information Sciences: Artificial intelligence, intelligent systems, multi agent systems
Gender: M

Years from PhD: 2

Prof Robert Wrembel (Poznan University of Technology [Faculty of Computing])
Participating as Secondary Proposer
E-mail: robert.wrembel@cs.put.poznan.pl
Telephone: +48 61 6653420
Core Expertise: Computer and Information Sciences: data warehouses, databases, data integration
Gender: M
Years from PhD: 17

Mr Jacek Kawalec (Voicelab.AI)
Participating as Secondary Proposer
E-mail: jacek@voicelab.ai
Telephone: +48 601173738
Core Expertise:
Gender: M
Years from PhD: 0

Portugal

Dr Sergio Matos (Universidade de Aveiro)
Participating as Secondary Proposer
E-mail: aleixomatos@ua.pt
Telephone: +351234370510
Core Expertise: Computer and Information Sciences: Text Mining
Gender: M
Years from PhD: 11

Prof Helena Galhardas (INESC-ID [IDSS Action Line])
Participating as Secondary Proposer
E-mail: helena.galhardas@tecnico.ulisboa.pt
Telephone: +351214233246
Core Expertise: Computer and Information Sciences: Databases, Data Cleaning and Integration
Gender: F
Years from PhD: 17

Romania

Prof Florin POP (University Politehnica of Bucharest)
Participating as Secondary Proposer
E-mail: florin.pop@cs.pub.ro
Telephone: +40723243958
Core Expertise: Computer and Information Sciences: Algorithms, distributed, parallel and network algorithms
Gender: M
Years from PhD: 10

Prof Dorian Gorgan (Technical University of Cluj-Napoca [Computer Science Department])
Participating as Secondary Proposer
E-mail: dorian.gorgan@cs.utcluj.ro
Telephone: +40264401478
Core Expertise: Electrical engineering, electronic engineering, Information engineering: Computer systems, parallel/distributed systems
Gender: M
Years from PhD: 24

Mr Ciprian-Octavian Truica (University POLITEHNICA of Bucharest [Computer Science and Engineering Department, Faculty of Automatic Control and Computers])

Participating as Secondary Proposer
E-mail: ciprian.truica@cs.pub.ro
Telephone: +40723285401
Core Expertise: Electrical engineering, electronic engineering, Information engineering: Databases, data mining, data curation, computational modelling
Gender: M
Years from PhD: 0

Prof Costin Badica (University of Craiova - Faculty of Automation, Computer and Electronics)
Participating as Secondary Proposer
E-mail: cbadica@software.ucv.ro
Telephone: +40251438198
Core Expertise: Computer and Information Sciences: Artificial intelligence, intelligent systems, multi agent systems
Gender: M
Years from PhD: 19

Ms Claudia Ifrim (Politehnica University Of Bucharest)
Participating as Secondary Proposer
E-mail: ifrim.claudia@gmail.com
Telephone: +40745940499
Core Expertise: Electrical engineering, electronic engineering, Information engineering: Development of scientific computing, data processing, simulation and modelling tools
Gender: F
Years from PhD: 0

Serbia

Dr RANKA STANKOVIC (University of Belgrade - Faculty of Mining and Geology)
Participating as Secondary Proposer
E-mail: ranka.stankovic@rgf.bg.ac.rs
Telephone: +381113219212
Core Expertise: Computer and Information Sciences: Theory of scientific computing and data processing
Gender: F
Years from PhD: 9

Prof Boris Delibasic (University of Belgrade - Faculty of Organisational Sciences)
Participating as Secondary Proposer
E-mail: boris.delibasic@fon.bg.ac.rs
Telephone: +381698893154
Core Expertise: Economics and business: Databases, data mining, data curation, computational modelling
Gender: M
Years from PhD: 11

Dr DRAGAN IVANOVIC (University of Novi Sad, Serbia - Faculty of Technical Sciences, Novi Sad)
Participating as Secondary Proposer
E-mail: chenejac@uns.ac.rs
Telephone: +381643560109
Core Expertise: Electrical engineering, electronic engineering, Information engineering: Software engineering, operating systems, computer languages
Gender: M
Years from PhD: 8

Prof MIRJANA IVANOVIC (University of Novi Sad - University of Novi Sad, Faculty of Sciences, Department of Mathematics and Informatics)
Participating as Secondary Proposer

E-mail: mira@dmi.uns.ac.rs
Telephone: +38121458888
Core Expertise: Computer and Information Sciences: Artificial intelligence, intelligent systems, multi agent systems
Gender: F
Years from PhD: 26

Slovakia

Dr Peter Butka (Technical University of Kosice - Faculty of Electrical Engineering and Informatics)
Participating as Secondary Proposer
E-mail: peter.butka@tuke.sk
Telephone: +421556024219
Core Expertise: Computer and Information Sciences: Theoretical aspects of data curation, data mining and database handling
Gender: M
Years from PhD: 8

Prof Maria Bielikova (Slovak University of Technology in Bratislava - Faculty of Informatics and Information Technologies)
Participating as Secondary Proposer
E-mail: maria.bielikova@stuba.sk
Telephone: +421902911888
Core Expertise: Electrical engineering, electronic engineering, Information engineering: Databases, data mining, data curation, computational modelling
Gender: F
Years from PhD: 23

Dr Martin Sarnovsky (Technical University Kosice - Faculty of electrical engineering and informatics)
Participating as Secondary Proposer
E-mail: martin.sarnovsky@tuke.sk
Telephone: +421556024219
Core Expertise: Computer and Information Sciences: Machine learning algorithms
Gender: M
Years from PhD: 9

Dr Giuseppe Lugano (University of Zilina [University Science Park])
Participating as Secondary Proposer
E-mail: giuseppe.lugano@uniza.sk
Telephone: +421944931080
Core Expertise: Media and communications: Media and communications, social aspects of information science and surveillance, socio-cultural communication
Gender: M
Years from PhD: 8

Dr Lubos Buzna (University of Zilina [University SCience Park])
Participating as Secondary Proposer
E-mail: Lubos.Buzna@fri.uniza.sk
Telephone: +421944278740
Core Expertise: Computer and Information Sciences: Mathematics applied to computer science, mathematical aspects of computer science
Gender: M
Years from PhD: 15

Slovenia

Dr Mirjana Kljajić Borštnar (University of Maribor - Faculty of Organizational Sciences)

Participating as Secondary Proposer
E-mail: mirjana.kljajic@fov.uni-mb.si
Telephone: +3862374236
Core Expertise: Other social sciences: Databases, data mining, data curation, computational modelling
Gender: F
Years from PhD: 12

Dr Marjan Čeh (University of Ljubljana)

Participating as Secondary Proposer
E-mail: marjan.ceh@fgg.uni-lj.si
Telephone: +38614768653
Core Expertise: Earth and related Environmental sciences: Databases, data mining, data curation, computational modelling
Gender: M
Years from PhD: 16

Mr Nejc Bat (Arctur d.o.o. [HPC & Cloud department])

Participating as Secondary Proposer
E-mail: nejc.bat@arctur.si
Telephone: 0038641400987
Core Expertise: Computer and Information Sciences: High Performance Computing
Gender: M
Years from PhD: 0

Mr Marko Praprotnik (XLAB d.o.o.)

Participating as Secondary Proposer
E-mail: marko.praprotnik@xlab.si
Telephone: +38612447782
Core Expertise:
Gender: M
Years from PhD: 0

Spain

Prof Antonio Fariña Martínez (University of A Coruña)

Participating as Secondary Proposer
E-mail: fari@udc.es
Telephone: +34981167000
Core Expertise: Computer and Information Sciences: Text compression and indexing. Compact data structures
Gender: M
Years from PhD: 13

Prof José Ramón Ríos Viqueira (Universidade de Santiago de Compostela)

Participating as Secondary Proposer
E-mail: jrr.viqueira@usc.es
Telephone: +34881816463
Core Expertise: Electrical engineering, electronic engineering, Information engineering: Databases, data mining, data curation, computational modelling
Gender: M
Years from PhD: 15

Dr Raquel Trillo Lado (Universidad de Zaragoza)

Participating as Secondary Proposer
E-mail: raqueltl@unizar.es
Telephone: +34976555539
Core Expertise: Electrical engineering, electronic engineering, Information engineering: Databases, data mining, data curation, computational modelling

Gender: F
Years from PhD: 6

Dr Miguel A. Martínez-Prieto (Universidad de Valladolid)
Participating as Secondary Proposer
E-mail: migumar2@infor.uva.es
Telephone: +34686117292
Core Expertise: Computer and Information Sciences: Theory of scientific computing and data processing
Gender: M
Years from PhD: 8

Dr SERGIO ILARRI (University of Zaragoza [Computer Science and Systems Engineering])
Participating as Secondary Proposer
E-mail: silarri@unizar.es
Telephone: +34876555262
Core Expertise: Electrical engineering, electronic engineering, Information engineering: Pervasive and ubiquitous computing
Gender: M
Years from PhD: 12

Dr Martín López Nores (Universidade de Vigo)
Participating as Secondary Proposer
E-mail: mlnores@det.uvigo.es
Telephone: +34625402112
Core Expertise: Computer and Information Sciences: Artificial intelligence, intelligent systems, multi agent systems
Gender: M
Years from PhD: 12

Prof David Camacho (Universidad Autonoma de Madrid)
Participating as Secondary Proposer
E-mail: david.camacho@uam.es
Telephone: +34914972288
Core Expertise: Computer and Information Sciences: Artificial intelligence, intelligent systems, multi agent systems
Gender: M
Years from PhD: 17

Dr Diego López-de-Ipiña (Universidad de la Iglesia de Deusto - DeustoTech -- Deusto Institute of Technology / University of Deusto)
Participating as Secondary Proposer
E-mail: dipina@gmail.com
Telephone: 619400104
Core Expertise: Computer and Information Sciences: Theoretical aspects of pervasive and ubiquitous computing
Gender: M
Years from PhD: 16

Dr José F. Aldana (University of Málaga [Ada Byron Research Center])
Participating as Secondary Proposer
E-mail: jfam@lcc.uma.es
Telephone: +34952132813
Core Expertise:
Gender: M
Years from PhD: 20

Dr Maria del Mar Roldan-Garcia (Universidad de Málaga [Centro de Investigación Ada Byron])

Participating as Secondary Proposer
E-mail: mmar@lcc.uma.es
Telephone: +34951952922
Core Expertise: Computer and Information Sciences: Semantic Web, Ontologies, Semantic Reasoning, Smart Data
Gender: F
Years from PhD: 6

Dr Ismael Navas-Delgado (Universidad de Málaga)
Participating as Secondary Proposer
E-mail: ismael@lcc.uma.es
Telephone: +34951952921
Core Expertise: Computer and Information Sciences: Data Management, Data Integration, Big Data, Linked Data
Gender: M
Years from PhD: 9

Prof Juan C. Trujillo (University of Alicante [Language and Information Systems])
Participating as Secondary Proposer
E-mail: jtrujillo@dlsi.ua.es
Telephone: +34965903400
Core Expertise: Electrical engineering, electronic engineering, Information engineering: Databases, data mining, data curation, computational modelling
Gender: M
Years from PhD: 17

Dr JAVIER NOGUERAS ISO (Universidad de Zaragoza [Advanced Information Systems Laboratory (IAAA)])
Participating as Secondary Proposer
E-mail: jnog@unizar.es
Telephone: +34876555533
Core Expertise: Computer and Information Sciences: Artificial intelligence, intelligent systems, multi agent systems
Gender: M
Years from PhD: 14

Dr Francisco J Lopez-Pellicer (Universidad Zaragoza [Advanced Information Systems Laboratory (IAAA), Aragon Institute for Engineering Research (I3A)])
Participating as Secondary Proposer
E-mail: fjlopez@unizar.es
Telephone: +34 876555552
Core Expertise: Electrical engineering, electronic engineering, Information engineering: Databases, data mining, data curation, computational modelling
Gender: M
Years from PhD: 7

Mr Roberto Garcia (enxenio sl)
Participating as Secondary Proposer
E-mail: rpazos@enxenio.es
Telephone: +34981913768
Core Expertise: Other engineering and technologies: Databases, data mining, data curation, computational modelling for other engineering and technologies
Gender: M
Years from PhD: 0

Mr Marcos Sacristan (TREELOGIC Telemática y Lógica Racional para la Empresa Europea S.L. [R&D])
Participating as Secondary Proposer

E-mail: marcos.sacristan@treelogic.com
Telephone: +34663246699
Core Expertise: Electrical engineering, electronic engineering, Information engineering: Computer systems, parallel/distributed systems
Gender: M
Years from PhD: 0

Mr Gabriel González (Deicom Technologies S. L.)
Participating as Secondary Proposer
E-mail: gabriel@deicom-technologies.com
Telephone: +34 886312510
Core Expertise: Electrical engineering, electronic engineering, Information engineering: Sensors and sensor systems
Gender: M
Years from PhD: 0

Switzerland

Prof Gilles Falquet (University of Geneva [CUI])
Participating as Secondary Proposer
E-mail: gilles.falquet@unige.ch
Telephone: +0041223790162
Core Expertise:
Gender: M
Years from PhD: 29

Prof Fabio Crestani (Universitá della Svizzera Italiana (USI) [Faculty of Informatics])
Participating as Secondary Proposer
E-mail: fabio.crestani@usi.ch
Telephone: +41586664657
Core Expertise: Computer and Information Sciences: Theoretical aspects of data curation, data mining and database handling
Gender: M
Years from PhD: 21

Prof Stephane Marchand-Maillet (University of Geneva [Department of Computer Science])
Participating as Secondary Proposer
E-mail: Stephane.Marchand-Maillet@unige.ch
Telephone: +41223790154
Core Expertise: Electrical engineering, electronic engineering, Information engineering: Databases, data mining, data curation, computational modelling
Gender: M
Years from PhD: 21

Turkey

Dr Burak Acar (Bogazici University)
Participating as Secondary Proposer
E-mail: acarbu@boun.edu.tr
Telephone: +902123596465
Core Expertise: Electrical engineering, electronic engineering, Information engineering: Signal/Image Processing and Analysis, Computer Aided Diagnosis/Prognosis, Medical Imaging
Gender: M
Years from PhD: 18

Mr Cem Çolak (Geovision Group [R&D Team])
Participating as Secondary Proposer
E-mail: cem.colak@gvg.com.tr
Telephone: +902123279205

Core Expertise: Electrical engineering, electronic engineering, Information engineering: Computer vision
Gender: M
Years from PhD: 0

United Kingdom

Dr Paolo Missier (Newcastle University)
Participating as Secondary Proposer
E-mail: Paolo.Missier@ncl.ac.uk
Telephone: +447891866167
Core Expertise: Computer and Information Sciences: Theory of scientific computing and data processing
Gender: M
Years from PhD: 11

Dr Andrea Calì (Birkbeck, University of London)
Participating as Secondary Proposer
E-mail: andrea@dcs.bbk.ac.uk
Telephone: +442076316796
Core Expertise: Computer and Information Sciences: Theoretical aspects of data curation, data mining and database handling
Gender: M
Years from PhD: 15

Dr Muhammad Ajmal Azad (Newcastle University)
Participating as Secondary Proposer
E-mail: majmalazad@gmail.com
Telephone: +351912691787__
Core Expertise: Computer and Information Sciences: Cryptology, security, privacy
Gender: M
Years from PhD: 2

Dr George Konstantinidis (University of Southampton [Electronics and Computer Science])
Participating as Secondary Proposer
E-mail: g.konstantinidis@soton.ac.uk
Telephone: +442380592978
Core Expertise: Computer and Information Sciences: Theoretical aspects of data curation, data mining and database handling
Gender: M
Years from PhD: 3

Dr Valentina Tamma (University of Liverpool [Department of Computer Science])
Participating as Secondary Proposer
E-mail: V.Tamma@liverpool.ac.uk
Telephone: +44 151 7954246
Core Expertise: Computer and Information Sciences: Artificial intelligence, intelligent systems, multi agent systems
Gender: F
Years from PhD: 16

Dr Thanassis Tiropanis (University of Southampton)
Participating as Secondary Proposer
E-mail: tt2@ecs.soton.ac.uk
Telephone: +442380599109
Core Expertise: Computer and Information Sciences: Web Science
Gender: M
Years from PhD: 17

Dr Adriane Chapman (University of Southampton [School of Electronics and Computer Science])

Participating as Secondary Proposer
E-mail: adriane.chapman@soton.ac.uk
Telephone: +44(0)23 8059 5000
Core Expertise: Computer and Information Sciences: Data Management
Gender: F
Years from PhD: 10

Dr Ioannis Komnios (EXUS Software Ltd [EXUS Innovation])

Participating as Secondary Proposer
E-mail: ikomnios@ee.duth.gr
Telephone: +306945406585
Core Expertise: Electrical engineering, electronic engineering, Information engineering: Networking
Gender: M
Years from PhD: 3

fYR Macedonia

Dr Gjorgji Madjarov (University Ss. Cyril and Methodius - Faculty of Computer Science and Engineering [Department of Software Engineering])

Participating as Secondary Proposer
E-mail: gjorgji.madjarov@finki.ukim.mk
Telephone: +38923099159
Core Expertise: Computer and Information Sciences: Machine learning algorithms
Gender: M
Years from PhD: 6

Dr Atanas Hristov (University of Information Science and Technology, Ohrid, Macedonia)

Participating as Secondary Proposer
E-mail: atanas.hristov@uist.edu.mk
Telephone: +38975229463
Core Expertise: Computer and Information Sciences: Algorithms, distributed, parallel and network algorithms
Gender: M
Years from PhD: 5