Data Management Research at the Technical University of Crete

Stavros Christodoulakis, Minos Garofalakis, Euripides G.M. Petrakis, Antonios Deligiannakis, Vasilis Samoladas, Ekaterini Ioannou, Odysseas Papapetrou, Stelios Sotiriadis
School of Electronic and Computer Engineering, Technical University of Crete

1. INTRODUCTION

The Technical University of Crete (TUC, www.tuc.gr) founded in 1977 in Chania, Crete is the youngest of the two technical universities in Greece (the other being the National Technical University of Athens). The purpose of this state institution is to provide high-quality undergraduate as well as graduate studies in modern engineering fields demanded by the Greek and international job market, to conduct research in cutting edge technologies as well as to develop links with the Greek and European industry. Today, the Technical University of Crete comprises five Engineering Schools (Electronic and Computer Engineering, Production Engineering and Management, Mineral Resources Engineering, Environmental Engineering, and Architecture). The School of Electronic and Computer Engineering (ECE) at TUC (www.ece.tuc.gr) has achieved an excellent reputation for its research and teaching internationally. The department accepts about 150 undergraduate students each year and employs 28 full-time faculty members. More than 75% of the ECE faculty members have obtained their Ph.D. degrees in top-rated foreign Universities, and several held academic or research positions abroad for many years prior to joining TUC. Faculty credits include multiple best paper awards at the ACM and IEEE Society level, professional recognition in terms of associate editor and technical committee member appointments, and leadership in conference organization here and abroad. Many TUC ECE graduates have pursued graduate studies at TUC and abroad. Their ranks include faculty members at top-rated North-American and European Universities, researchers at University, Government, and Industrial Research Labs, and successful professional engineers across Greece and Europe.

Data-management research at TUC revolves around a broad and diverse range of topics, ranging from fundamental algorithmic techniques (e.g., for managing streaming and probabilistic data) and tools for big-data analytics, to cloud database architectures, digital libraries, and the semantic web. In this short article, we present an overview of some recent and ongoing data-management research efforts at TUC-ECE. We structure our discussion by grouping our research activities under each of the three main data-management research labs at TUC-ECE: The Software Technology and Network Applications (SoftNet) Lab (www.softnet.tuc.gr, headed by Prof. Minos Garofalakis), the Intelligent Systems Lab (www.intelligence.tuc.gr, headed by Prof. Euripides G.M. Petrakis), and the Distributed Multimedia Information Systems and Applications (MUSIC) Lab (www.music.tuc.gr, headed by Prof. Stavros Christodoulakis).

2. SOFTNET LAB

Continuous Monitoring of Distributed Streaming Data. Large-scale stream processing applications rely on continuous, event-driven monitoring, that is, real-time tracking of measurements and events, rather than one-shot answers to sporadic queries. Furthermore, the vast majority of these applications are inherently distributed, with several remote monitor sites observing their local, high-speed data streams and exchanging information over a communication network. This distribution of the data naturally implies critical communication constraints that typically prohibit centralizing all the streaming data, due to either the huge volume of the data (e.g., in IP-network monitoring), or power and bandwidth restrictions (e.g., in wireless sensornets). Finally, an important requirement of large-scale event monitoring is the effective support for tracking complex, holistic queries that provide a global view of the data by combining and correlating information across the collection of remote monitor sites. Monitoring the precise value of such holistic queries without continuously centralizing all the data at first seems hopeless. Given the prohibitive cost of data centralization, it is clear that realizing sophisticated, large-scale distributed data-stream analysis tools must rely on novel algorithmic paradigms for processing local streams of data *in situ* (i.e., locally at the sites where the data is observed). This, of course, implies the need for intelligently decomposing a (possibly complex) global data-analysis and monitoring query into a collection of "safe" local queries that can be tracked independently at each

site (without communication), while guaranteeing correctness for the global monitoring operation. This decomposition process can enable truly distributed, event-driven processing of real-time streaming data, using a push-based paradigm, where sites monitor their local queries and communicate only when some local query constraints are violated. Nevertheless, effectively decomposing a complex, holistic query over the global collections of streams into such local constraints is far from straightforward, especially in the case of non-linear queries (e.g., joins).

A useful tool for monitoring complex non-linear queries over distributed streams is the recently proposed geometric approach [15, 8]. In a nutshell, the geometric method enables the monitoring of complex non-linear functions expressed over the *average* of data vectors maintained at distributed sites. The monitoring is made possible by having each site monitor a geometric condition over the *domain* where the average vector lies, rather than monitoring the range of the function. These local geometric constraints are designed to guarantee that, if the monitored condition on the global function is violated, then at least one of the local constraints must be violated, that is, at least one of the remote sites will fire a *local violation*. Thus, no global violation can go undetected.

Our recent work, in the context of the LIFT EU-FET Open project (`www.lift-eu.org`), builds on the geometric framework in order to solve a variety of complex distributed stream monitoring problems, including: Detecting outliers in sensor networks by monitoring the pair-wise similarities (which can be expressed as a wide range of functions, including, for example, L_k norms, cosine similarity, and Extended Jaccard coefficient) of neighboring sensor nodes [3]; efficiently monitoring complex functions by combining the use of *prediction models* with the geometric approach [6]; monitoring sliding-window queries by efficiently summarizing streaming data over sliding windows with probabilistic accuracy guarantees [14]; enriching the geometric approach with *sketch synopses* [4] to efficiently track a broad class of complex queries (including, general inner products, self-join sizes and range aggregates) over massive, high-dimensional distributed data streams with provable guarantees [5]; monitoring continuous fragmented skyline queries over distributed data streams [13]; and, proposing novel techniques for defining improved *safe zones* (i.e., safe regions of the domain for local data vectors) for distributed monitoring problems [7]

Our ongoing work in the area of (centralized and distributed) data-stream management focuses on novel extensions of the technology and tools to handle the challenges of (1) large-scale Complex Event Processing (CEP) systems (in the context of the upcoming FERARI EU-STREP project), (2) massive brain data analytics (in the context of the EU FET-Flagship Human Brain Project, `www.humanbrainproject.eu`), and (3) new, elastic software/hardware architectures for effective data-stream analytics (in the context of the upcoming QualiMaster EU-STREP project).

Data-as-a-Service (DaaS) in Microcloud Federations. Collecting, storing, and processing public web-size data, such as the web graph and public data from social networks, has for long being an exclusive privilege of a few large companies world-wide that have the capacity to construct and maintain huge server farms. Towards enabling small and medium companies to perform management and mining tasks on data of such magnitude, we have recently, in the context of the LEADS EU-STREP project (`www.leads-project.eu`), started exploring an innovative cloud model, called Data-as-a-Service (DaaS). The model enables companies to use shared cloud resources for storing and accessing public and private data, and for performing arbitrary processing tasks on this data. The targeted infrastructure in our case is an elastic set of distributed microclouds, combined to create the illusion of a large unified cloud.

The considered scenario has several key benefits compared to traditional in-house solutions. First and foremost, companies can share the acquisition (e.g., crawling) and storage cost of the public data. Results of common processing tasks, such as the PageRank scores of web-pages or the influence factors of users in social networks, can also be shared across platform users. Second, companies can use a pay-as-you-go charge model, without requiring upfront investment. This enables small companies to test innovative, high-risk, ideas, without a substantial investment. Last, sharing of the infrastructure reduces the idle time of the participating nodes, promoting green computing and reducing the platform's running cost.

The model also comes with a novel set of challenges. Probably the key concerns for companies are the correctness of the data and results, and the privacy of sensitive data. Therefore, in a recent paper we have considered the problem of verifying the correctness and freshness of query results on data streams, necessary in the existence of malicious or misbehaving nodes in the network [12]. Our solution induces a very small overhead, and is readily applicable to generic cloud setups. In the same context, we recently started investigating the problem of data analytics on private, encrypted, data in the cloud. Our recent results (working paper) show that many of the queries necessary for powerful data analytics can in fact be executed without information leakage, directly in the cloud.

The physical distribution of the individual clouds in the considered infrastructure offers many optimization

opportunities. Data is partitioned on servers distributed across the world, each one with different (possibly even varying) computational capacity and charging policies. By controlling the data placement and replication, the location of the processing tasks, as well as the network interaction between the nodes, we can drastically reduce the running cost of the platform. Keeping this in mind, we are now developing novel distributed algorithms for frequent data processing tasks over both streaming and stored data, which rely on approximation and on in situ processing. Preliminary results on maintenance of PageRank scores for the web graph show that these techniques substantially reduce the cost without noticeable impact on quality (working paper).

Uncertain/Probabilistic Data Management. We are also working on developing techniques for efficient management of uncertain data, originating, for example, from information extraction and resolution processes. The majority of the work on this topic is done in the context of the HeisenData project (`heisendata.softnet.tuc.gr`), which aims at extending the traditional relational table store with support for a broad class of statistical models and probabilistic-reasoning tools.

Part of our research focus involves efficient query processing over data extracted from unstructured sources [23, 22, 24]. For instance, the possible extractions can be represented using the Conditional Random Field (CRF) statistical model [18], and inference over such a CRF provides the final extraction results. The system presented in [24], is a probabilistic framework that allows performing such extraction tasks. It uses a linear-chain CRF with the Viterbi inference algorithm and query processing returns the maximum-likelihood extraction results. The in-database implementation of extraction tasks, as introduced in [18], improves the quality of query results as well as the efficiency of query processing since it enables the incorporation of several optimizations. The approach in [23] considers additional inference algorithms, such as variations of the sampling-based Markov chain Monte Carlo. It also proposes mechanisms for choosing the most suitable inference algorithm based on the given data, model, and query. We are currently working on further improving quality and efficiency by combining additional extraction and database activities, including coreference, canonicalization and optimizations using algebraic equivalences. Providing such a system requires addressing several challenges, such as efficiently executing the inference process on the potentially large graphical models that will be created. To enable large-scale probabilistic data analysis, we have recently developed a novel MapReduce algorithm. It efficiently executes exact inference on large graphs by taking advantage of the parallel nature of exact inference both structurally and computationally.

With respect to management of data from entity resolution methods, we are considering probabilistic unmerged duplicates specifying which objects can be merged. More specifically, we proposed an entity-join operator that allows expressing joins between the tables containing unmerged duplicates with other tables from the database. The focus is on analytics that allow users to express aggregation and iceberg queries over the massive collection of "possible resolution worlds". Processing is based on a novel indexing structure that allows efficient access to the resolution-related information and a set of techniques for evaluating complex queries that include qualifiers for retrieving analytical and summarized information and moving towards a higher level of detail. An extension of this work is to reduce the time required for query processing by considering also a set of apriori merges along with the on-the-fly merges created during query processing.

3. INTELLIGENT SYSTEMS LAB

During the latest years we have witnessed the rapid growth of cloud computing that delivers leased services to everyday Internet users. Various provisioning models have been defined to separate accessibility and content of services. A fundamental taxonomy includes infrastructural services (Infrastructure as a Service-IaaS), software based services (Software as a Service-SaaS) and development platforms (Platform as a Service-PaaS). Initially, IaaS is related to active and virtualized services that scale dynamically while SaaS refers to applications that are already hosted in cloud datacenters. At last, PaaS could be seen as an outgrowth of SaaS, wherein applications are available in a development platform environment where users could implement their own cloud based solutions. Cloud computing has been proven to be a novel approach, for many cases, regarding to the minimization of operational and monitoring costs while it increases elasticity. However in the healthcare domain, in particular, there are standards, regulations and recommendations (e.g., national legislation, ISOs and security standards). These stress severe restrictions for data transfer, storage, aggregation and analysis. For the case of cloud computing a typical requirement is that services presumed to be stored in remote datacenters, while the data storage happens, as well, remotely. This raises obstacles and the utilization of cloud capabilities by healthcare domain seems challenging.

We are taking part in the Future Internet Public-Private Partnership (FI-PPP, `www.fi-ppp.eu`) of the EU (a European programme for Internet-enabled innovation). The ongoing phase 2 of FI-PPP includes 5 use case trials, one of which is the FI-Star project (`www.fi-star.eu`) that aims at establishing early trials in the Health Care domain building on Future Internet (FI) technology lever-

aging on the outcomes of FI-PPP Phase 1. FI-STAR adopts a fundamentally different, "reverse" cloud approach that is bringing the software to the data, rather than bringing the data to the software [19] as a means of developing Future Internet (FI) applications. At a glance, cloud services (SaaS) could reach the local user infrastructure and utilized in an on demand model while services are deployed locally. This highlights new challenges in the area of designing FI applications for sensitive domains like the e-health domain where data security and confidentiality raise obstacles on data processing (e.g., data may not migrate to a public cloud and need to be processed locally).

Within the FI-Star project, we focus on exploiting a SaaS cloud model for deploying an architecture solution [17] that resorts to FI-WARE . FI-WARE is a core software platform that eases creation of innovative Future Internet (FI) applications by offering reusable modules named as Generic Enablers (GEs). Generic Enablers (GEs) are considered as software units (the core building blocks of a cloud application) that offer various functionalities along with protocols and interfaces for operation and communication.

Particular emphasis is given to interoperability and portability of cloud services. This drives from the need of translating services across different cloud providers and the need for service discovery within each cloud platform. Thus, providers could allow services to communicate and interoperate. Service interoperability refers to the interoperation of services across multiple clouds using a common management API while, system portability defines the ability of cloud services to be deployed on other cloud service of a lower service model (for instance a SaaS to be integrated in a PaaS). Our work is also closely related with the case of portability wherein an application needs to be ported in the cloud, and this is particularly interesting for utilizing the legacy applications and systems. We are motivated from the exploration of semantic annotated descriptions in order to assist cloud porting in terms of service automate operations as well as assign rich content and relationships. Practically, this requires definition of the service (for instance using service descriptions) in order to be translated and be compatible with a cloud.

One of our research aims is to explore ontology-based solutions, for representing cloud services specific concepts, attributes and relationships. In recent work we have presented an analysis to define the requirements for interoperability and portability in various cloud deployment models. A future task is to design service descriptions of GEs and automate the service discovery process. This will offer significant advantages for locating appropriate GEs that integrate FI applications.

Recent work includes developments in our lab's state-of-the-art cloud setting (termed Intellicloud, `cloud.intellicloud.tuc.gr`) that will offer several kinds of provider services for FI application development including GEs. Intellicloud is based on Open Cloud Computing Interface (OCCI, `occi-wg.org`) standard that is an open specification and API for cloud offerings and it is aligned with FI-WARE. OCCI promotes developments in the area of interoperability and portability, areas that we perform our research. Currently, Intellicloud offers IaaS and PaaS services (that integrate GEs) to researchers for experimentation and implementation of FI applications (e.g., healthcare services)

In many instances, healthcare services have been developed based on the IoT paradigm that enables devices to be represented in an Internet like structure. In FI-STAR, provider cloud services will manage and upgrade functionalities of GEs and will be deployed in a customize way that matched the health care use case constraints. Fundamentally, this includes the cloud management for supervision, underlying infrastructure, the utilization of various IoT devices for data collection, provision of APIs (e.g., tools for data analytics) and communication among interfaces. This in combination with edge computing expands clouds functionality by allowing business logic and process management to happen at the actual source similar to a distributed computing fashion. This characterizes an alternative view of clouds where services can reach user premises and utilized directly in users personal IoT devices. We also focus on the exploration of healthcare provisioning models and approaches. For instance we have designed decision support system for patients suffering from Bipolar Disorder. Thus, the adaptation of performed work by utilizing emerging technologies highlights an area of our new challenges.

We vision that edge computing could offer cloud capabilities for remote data storage and management, while local data processing will facilitate a self-adaptive environment for data extraction and analysis. In such solution, legacy or on-the-fly developments will need to be imported to the cloud infrastructure and to interoperate in both local and remote clouds. This means that users software and APIs will need to communicate successfully and understand the new system constraints, a case that highlights cloud portability.

4. MUSIC LAB

Semantic Web Interoperability. The dominant standard for information exchange in the web today is XML and many important international standard have been expressed in XML and its derivatives. The emerging Semantic Web (SW) world however is based on different models and languages. We investigate methodologies for bridging the gap between the Semantic Web and the

XML world. A survey and comparison of recent technologies and standards in the XML and SW world as well as the data integration and data exchange systems between the two appears in [1]. We have developed the XS2OWL Framework ([1]) which provides a transformation model for automatic and accurate expression of the XML Schema Semantics in OWL Syntax. In addition it allows it allows the transformation of the XML data in RDF format and vice versa. The current version of XS2OWL exploits the OWL 2 semantics (like identity constraints) and supports the new XML constructs introduced in XML Schema 1.1. We have also developed the SPARQL2XQuery Framework for the translation of SPARQL to XQuery [2]. Although several systems offer SPARQL end points over relational data there are no systems supporting XML data. The Framework provides a formal model for the expression of mappings from OWL to XML Schema and a generic method for SPARQL to XQuery translation, thus providing an important part of ontology based integration involving XML resources in the Linked data environment.

Driven by the fact that Semantic Web is comprised of distributed, diverse (in terms of schema adopted) and in some cases overlapping RDF datasets, we are investigating generic frameworks supporting query answering in federated architectures. To this end, the SPARQL–RW Framework [9] provides a formal method and implementation for SPARQL query rewriting with respect to a set of predefined mappings between ontology schemas. The supported mapping model has been formally described using Description Logics and allows the definition of a rich set of mapping types. Our Framework is proved to provide semantics preserving queries to the nodes.

Finally, important international standards such as MPEG-7 for multimedia content descriptions do not describe a formal mechanism for the systematic integration of domain knowledge in the MPEG-7 descriptions. We have developed a formal model for domain knowledge representation within MPEG-7 [20]. The model allows the systematic integration of domain knowledge in MPEG-7 descriptions using only MPEG-7 constructs thus maintaining interoperability with existing MPEG-7 software.

Meta Data Management, Interoperability and Linked Data Publishing in Museum Digital Libraries. In addition to web presence today museums are also very interested in interoperability with generic or domain specific large international metadata publishers as well as publishing their data in the semantic web world. In the context of the Natural Europe project we have developed methodologies, software architectures and systems to support Natural history museums for their web presence, their interoperability with major international metadata providers and search engines and for publishing their data as Linked Data ([10], [16]). The systems have been installed and used in six important European Natural History museums.

Each museum node is provided with a Multimedia Authoring Tool (MMAT), a Cultural heritage Object (CHO) repository and with a Vocabulary Server facilitating the complete metadata management lifecycle ingestion, maintenance, curation and dissemination of CHOs. The infrastructure also supports the migration of legacy metadata migration into the node. The application profile of CHOs has been created through an iterative process with the museum domain experts and it is a superset of the Europeana Semantic Elements (ESE) metadata format, thus providing a direct interoperability with the central European Digital library (Europeana). The CHO repository manages both content and metadata and adopts the OAIS Reference Model for ingestion, maintenance and dissemination of information packages. The Vocabulary Server supports any taxonomic classification that the museum may use. The ingested taxonomies follow the SKOS format which is a leading international standard based on the Semantic Web principles for representations of Thesauri, Taxonomies and other types of controlled vocabularies. The controlled vocabularies provide strong support for the curation, indexing, retrieval, autocomplete functionality, etc.

For Natural history an important vocabulary is is the Catalogue of life (CoL) which contains 1.4 millions of species and their relationships. We have expressed the taxonomy of CoL to SKOS using the CoL annual checklist and a D2R server. An Access Module provides a number of services that allow selective harvesting of metadata from external entities through an OAI-PMH interface. The museums can be seen individually or through a federation. The Access Module is used to harvest the metadata to the federal node, to Europeana, as well as to establishing connections with major biodiversity networks such as GBIF and BIOCASE. The BIOCASE network is based on a very involved Schema (ABCD Schema) which describes nearly 1200 different concepts. We developed in cooperation with museum experts mappings between the ABCD schema concepts, and wrappers to be used by BIOCASE to access the XML databases of natural Europe (the BIOCASE wrappers assume relational dbms underneath). The wrappers developed follow a layered architecture so that they can be easily adapted for other XML data sources.

To support the Semantic Web presence of each museum individually we have described in OWL the CHO Application profile of Natural Europe. The resulting Natural Europe ontology references well known ontologies/schemas (like DC, FOAF, Geonames, SKOS) and has been aligned with the Europeana Data model (EDM) supporting interoperability with the Europeana Seman-

tic Layer. The publication process involves establishing links to the external RDF data sets, conversion of the XML data to RDF, maintenance, publishing and dissemination of RDF data. The Semantic Infrastructure allows highly expressive queries combining knowledge from distributed resources like 'find photos of endangered species of genus "Bufo" in neighboring countries of Greece' which combines information from Natural Europe, DBpedia, CoL/Uniprot, and Geonames.

Management of Mobile Multimedia Nature Observations using Crowd Sourcing. Several scientific fields (biodiversity, biology, agriculture, etc.) would greatly benefit if informed users with interest in the domain could contribute with their observations to the knowledge in the domain. This need arises from the fact that the number of scientists and the available funding in certain domains are very limited with respect to the real needs, We are developing a Software Framework [21] that supports communities with common interests in nature to capture and share multimedia observations of nature objects or events using mobile devices. The observations are automatically associated with contextual objects (such as GPS objects, pictures, sensor data) and they can be visualized in a faceted manner on top of 2D or 3D maps. The observations are managed by a multimedia management system, and annotated by the same and/or other users with common interests. Multimedia observations of nature objects or events can be annotated by multimedia annotations that are complex resources. Annotations made by the crowd support the knowledge distillation and data provenance.

Collaborative Environments for Instructional Design. We are investigating the design of collaborative environments that allow instructional designers and educators to develop educational templates and scenarios that can be used in different educational contexts. We have developed such a tool, Octopus [11], in the context of EU projects. Octopus is compatible with IMS LD Level A, while hiding its complexity from its user interfaces. It supports a wide range of collaboration and interoperability features, and extensive usability tests have been used to improve its interfaces.

5. ADDITIONAL AUTHORS

Additional authors: N. Giatrakos, N. Gioldasis, P. Arapi, N. Moumoutzis, K. Stravoskoufos, A. Preventis, C. Tsinaraki, K. Makris, G. Skevakis, M. Mylonakis, I. Trohatou, V. Kalokyri, and V. Vazaios.

6. REFERENCES

[1] N. Bikakis, C. Tsinaraki, N. Gioldasis, I. Stavrakantonakis, and S. Christodoulakis. The XML and Semantic Web Worlds: Technologies, Interoperability and Integration: A Survey of the State of the Art. In *Semantic Hyper/Multimedia Adaptation: Schemes and Applications*. 2012.

[2] N. Bikakis, C. Tsinaraki, I. Stavrakantonakis, N. Gioldasis, and S. Christodoulakis. The SPARQL2XQuery Interoperability Framework. *WWW Journal*, 2013.

[3] S. Burdakis and A. Deligiannakis. Detecting outliers in sensor networks using the geometric approach. In *IEEE ICDE*, 2012.

[4] G. Cormode, M. Garofalakis, P. J. Haas, and C. M. Jermaine. Synopses for massive data: Samples, histograms, wavelets, sketches. *Foundations and Trends in Databases*, 4(1-3), 2012.

[5] M. Garofalakis, D. Keren, and V. Samoladas. Sketch-based geometric monitoring of distributed stream queries. *PVLDB*, 6(10), 2013.

[6] N. Giatrakos, A. Deligiannakis, M. N. Garofalakis, I. Sharfman, and A. Schuster. Prediction-based geometric monitoring over distributed data streams. In *ACM SIGMOD*, 2012.

[7] D. Keren, G. Sagy, A. Abboud, D. Ben-David, A. Schuster, I. Sharfman, and A. Deligiannakis. Geometric monitoring of heterogeneous streams. *IEEE TKDE*, 2013.

[8] D. Keren, I. Sharfman, A. Schuster, and A. Livne. Shape sensitive geometric monitoring. *IEEE TKDE*, 24(8), 2012.

[9] K. Makris, N. Bikakis, N. Gioldasis, and S. Christodoulakis. SPARQL-RW: Transparent Query Access over Mapped RDF Data Sources. In *EDBT*, 2012.

[10] K. Makris, G. Skevakis, V. Kalokyri, P. Arapi, and S. Christodoulakis. Metadata Management and Interoperability Support for Natural History Museums. In *TPDL*, 2013.

[11] M. Mylonakis, P. Arapi, N. Moumoutzis, S. Christodoulakis, and M. Ampatzaki. Octopus: A Collaborative Environment Supporting the Development of Effective Instructional Design. In *ICEEE*, 2013.

[12] S. Papadopoulos, G. Cormode, A. Deligiannakis, and M. Garofalakis. Lightweight authentication of linear algebraic queries on data streams. In *ACM SIGMOD*, 2013.

[13] O. Papapetrou and M. Garofalakis. Continuous fragmented skylines over distributed streams. In *IEEE ICDE*, 2014.

[14] O. Papapetrou, M. N. Garofalakis, and A. Deligiannakis. Sketch-based querying of distributed sliding-window data streams. *PVLDB*, 5(10), 2012.

[15] I. Sharfman, A. Schuster, and D. Keren. A geometric approach to monitoring threshold functions over distributed data streams. *ACM TODS*, 32(4), 2007.

[16] G. Skevakis, K. Makris, P. Arapi, and S. Christodoulakis. Elevating Natural History Museums' Cultural Collections to the Linked Data Cloud. In *SDA*, 2013.

[17] S. Sotiriadis, G. Petrakis, S. Covaci, P. Zampognaro, E. Georga, and C. Thuemmler. An architecture for designing Future Internet (FI) applications in sensitive domains: Expressing the Software to data paradigm by utilizing hybrid cloud technology. In *IEEE BIBE*, 2013.

[18] C. A. Sutton and A. McCallum. An introduction to conditional random fields. *Foundations and Trends in Machine Learning*, 4(4), 2012.

[19] C. Thuemmler, J. Mueller, S. Covaci, T. Magedanz, S. de Panfilis, T. Jell, A. Schneider, and A. Gavras. Applying the Software-to-Data Paradigm in Next Generation E-Health Hybrid Clouds. In *ITNG*, 2013.

[20] C. Tsinaraki and S. Christodoulakis. Domain Knowledge Representation in Semantic MPEG-7 Descriptions. In *The Handbook of MPEG applications: Standards in Practice*.

[21] C. Tsinaraki, G. Skevakis, I. Trochatou, and S. Christodoulakis. MoM-NOCS: Management of Mobile Multimedia Nature Observations using Crowd Sourcing. In *MoMM*, 2013.

[22] D. Z. Wang, M. J. Franklin, M. N. Garofalakis, and J. M. Hellerstein. Querying probabilistic information extraction. *PVLDB*, 3(1), 2010.

[23] D. Z. Wang, M. J. Franklin, M. N. Garofalakis, J. M. Hellerstein, and M. L. Wick. Hybrid in-database inference for declarative information extraction. In *ACM SIGMOD*, 2011.

[24] D. Z. Wang, E. Michelakis, M. J. Franklin, M. N. Garofalakis, and J. M. Hellerstein. Probabilistic declarative information extraction. In *IEEE ICDE*, 2010.