# Chapter 16
# A Cloud-Based Data Network Approach for Translational Cancer Research

**Wei Xing, Dimitrios Tsoumakos, and Moustafa Ghanem**

**Abstract** We develop a new model and associated technology for constructing and managing self-organizing data to support translational cancer research studies. We employ a semantic content network approach to address the challenges of managing cancer research data. Such data is heterogeneous, large, decentralized, growing and continually being updated. Moreover, the data originates from different information sources that may be partially overlapping, creating redundancies as well as contradictions and inconsistencies. Building on the advantages of elasticity of cloud computing, we deploy the cancer data networks on top of the CELAR Cloud platform to enable more effective processing and analysis of Big cancer data.

## 16.1 Introduction

Translational cancer research builds on incorporating multiple levels of biological information within clinical data with aim of gaining better understanding of how cancer works and developing new ways for identifying, preventing and treating the disease. The first challenge in conducting translational studies is that the data is heterogeneous, including phenotype, genotype, expression profiling, proteomics, protein interaction, metabolic analysis data as well as physiological measurements, etc. The second key challenge is that the data is large, decentralized, growing and continually being updated. It originates from different sources, e.g. different

W. Xing (✉)
Cancer Research UK Manchester Institute, University of Manchester,
Manchester M20 4BX, UK
e-mail: wei.xing@cruk.manchester.ac.uk

D. Tsoumakos
Computing Systems Laboratory, National Technical University of Athens,
Athens 15773, Greece
e-mail: dtsouma@cslab.ece.ntua.gr

M. Ghanem
School of Science and Technology, Middlesex University, Hendon, London NW4 4BT, UK
e-mail: m.ghanem@mdx.ac.uk

P. Vlamos, A. Alexiou (eds.), *GeNeDis 2014*, Advances in Experimental Medicine
and Biology 820, DOI 10.1007/978-3-319-09012-2_16

experiments and different labs. It is also stored in different distributed databases that are managed by different organizations. The content of such databases is thus typically overlapping, creating redundancies as well as contradictions and inconsistencies.

In this paper, we develop a new cancer data network (CDN) model and associated technology for constructing and managing **self-organizing** content in order to support the integration of biological and clinical data with the research it is spawned from. In addition, CDN offers the ability of track several aspects of patient care according to genetic and molecular profiles to facilitate tailoring of treatment.

The proposed CDN architecture is based on a novel content management model that supports end users in managing distributed, dynamic and evolving data sets. Within CDN, we shift the view of **content** from being a static resource, and introduce it as a **dynamic and intelligent entity** that is able to perform operations such as linking itself to other relevant content. In doing so, the intelligent content can discover implied relationships with other content, identifying redundancies and overlap as well as updating its links with the ecosystem when new content is added or old content is removed or depreciated.

Within the CDN approach, data content is represented as an active object equipped semantic mechanisms that allow a greater degree of flexibility towards automating the procedure of content management and organization. To this end, we define **Active Cancer Data Content** as a logical container that contains the digital data content (i.e., patient data, clinical data, research experiment data, publications, public gene or protein databases, etc.) together with intelligent and autonomic, self-organizing mechanisms for automating content management.

The remainder of this paper is organized as follows. Section 16.2 introduces the underlying principles and design of the CDN; Sect. 16.3 presents the architecture of CDN, focusing on its software components, and on the main interactions between them, and describing how each of them is instantiated for the implementation with EU CELAR cloud platform; Sect. 16.4 describes the related work; and finally, "Conclusion" section provides conclusions, and describes open issues and our planned future work.

## 16.2 The Design of CDN

CDN is designed to bridge the gap between translational research and targeted patient treatment. Hence, the design goals of CDN are (1) to better analyse data obtained dynamically from various bio-instrument sources in order to answer biological question at a system level; and (2) to better translate data obtained from in vitro and in vivo discoveries into the clinic.

## *16.2.1 Problems and Requirements*

The key challenge CDN addresses is that information stored or published over the web and other specialized data sources is heterogeneous, decentralized, growing and continually being updated. Furthermore, the content living or archiving on different information sources partially overlaps creating redundancies as well as contradictions and inconsistencies. In this section we describe the current issues of the area of personalized medicine research.

### 16.2.1.1 Redundant or Irrelevant Information of Protein and Gene Sequence

Currently, well over a thousand accessible data sources provide information pertaining to any gene, mRNA or protein sequence (estimated by the number of known SRS "Sequence Retrieval System") such as polymorphisms, protein interactions and expression levels. The vast majority of the data sources are specialized and are therefore maintained and updated by different organizations. In addition, data sources with the same emphasis (such as nucleotide or protein sequences) are updated and curated at different intervals and with various benchmarks and standards. As a result, many databases contain outdated, redundant or irrelevant information pertaining to the scientific question at hand.

Also, our continually expanding knowledge base adds new dimensions to the content. For example, the cataloging and assessment of functional impact of recently discovered mechanisms of dynamic biological regulation, including but not restricted to microRNAs and our knowledge of protein modification types and permutations, is incomplete. New categorical discoveries and their related information detail will need to be progressively built into any comprehensive content structure.

### 16.2.1.2 Evolving Methods of Data Generation from Multiple Instrument Platforms

Translational cancer research requires the integration of data from state-of-the-art technologies, for which the methods of translating and interpretating raw instrument data into relevant contextualized biological outputs are continually improving. An example of this is the interpretation of mass spectrometry peptide fragmentation data into qualitative and quantitative peptide and protein data in proteomics experiments. Different instruments produce data with different technical characteristics, including signal-to-noise ratios, raw signal intensities, and data accuracy, precision and resolution. These characteristics are continually changing for the better, but will continue to vary depending on the type and generation of instruments used, new hardware innovations, and the data acquisition and experiment style.

The bioinformatic translation of the raw fragmentation data into peptide and protein identities is also evolving. Current strategies typically employ probabilistic, stochastic or descriptive models to pattern match fragment ion profiles against theoretical profiles generated against assumed protein sequences and modification content. Personalized medicine will dictate a drift away from this data interrogation strategy since each individual labours genomic and proteomic differences that would not be represented in an assumed protein sequence database. This may involve fundamental changes to the data interrogation strategy, for example a migration towards de novo sequencing tools, or at the very least changes to scoring of genepeptideprotein sequence assignments and the specific identification of mutations, polymorphisms or variables specific to individuals.

#### 16.2.1.3 Creating Genomic Networks

To elucidate the wiring of cellular information processes, current research requires integration of quantitative and dynamic data from several sources. Such information sources could be genomic public database-based, sequence-based or clinical information-based and require various algorithms and software package for data analysis. For maximal output from such data, it is important that the multidimensionality is taken into account and the data can be visualized with differential weighting of individual data sources. For example, gene mutation and gene function interactions can be measured in a static manner using techniques such as yeast 2 hybrid and complement assays as well as the dynamic and quantitative abundances can be included through platforms such as COSMIC, VerScan and Meerkat runs. While each source provides important information, the sources provide complementary aspects of information, which is important to integrate and visualize.

To address the above issues, we design CDN system to support:

1. **Integrating heterogeneous and unstructured content.** It allows scientists to incorporate multiple levels of background information within their studies, such as phenotype, genotype, expression profiling, proteomics, protein–protein interactions, biochemical metabolic studies, and physiology measurement, etc.
2. **Decentralized control and collaborative communities.** The content itself either arises from biological experiments conducted by individual groups or as a result of data integration and analysis studies using data published by other groups.
3. **Multi-discipline.** The information is highly relevant to researchers working on other topics can be shared easily, between specialized data sources (including scientific literature) and databases focusing on specific topics, e.g. organisms, diseases, genes, proteins, metabolic pathways, chemical compounds or on relationships between them.

Our special focus is on addressing the issues of overwhelming and continuous flood of complex information generated and published on a daily basis through the use of Semantic Web Technology. We illustrate our approach in the next section.

## *16.2.2 Semantic Approach*

CDN aim to develop novel mechanisms for constructing and generating symbiotic, semantically described, self-aware and self-organizing content technology that enables distributed digital objects to be linked together into **CDNs**.

### 16.2.2.1 The CDNs

A key feature of scientific information is that it is continually evolving. For example, new information about scientific entities (e.g. proteins, genes or diseases) is being published on a daily basis. Furthermore, the decentralized authority over the content, whereby scientists in different organizations publish and manage their own findings, means that information about the same, similar or related entities, may be stored on different sources that evolve in different ways. This inevitably results in partial overlaps in the coverage of the data sources creating redundancies as well as contradictions and inconsistencies at both the entity and the concept level.

We design CDN to link individual elements of the digital content together. By using semantic data model and ontology, we define two types of links among the CDN nodes (i.e. content): explicit links and conceptual links.

Explicit links between different elements are typically stored with the content. At the simplest level an entry on a specific protein on a particular data source can make explicit references to other protein, gene or disease entries on other sources, or to specific supporting scientific publications. Ontologies can be used to either manually or automatically assign scientific papers, genes, proteins to different categories.

Conceptual links between different elements are typically not stored with the content, but can traditionally be inferred by using either statistical/probabilistic analysis techniques or domain knowledge. At the simplest level, users may wish to group proteins together based on the similarity of specific properties such as their effect on the same cellular function, or their causal implication to a similar disease phenotype.

### 16.2.2.2 Retrieval, Integration and Update

In [1], we developed a semantic information integration approach to integrate and update information from distributed, heterogeneous data sources dynamically.

CDN employs ActOn as a means to retrieve, integrate and manage the CDN Content in an intelligent, active manner.

ActOn is an ontology-based information integration approach that is suitable for highly dynamic distributed resources. To deal with this issue, i.e. that information changes frequently and information requests have to be answered quickly in order to provide up-to-date information, ActOn employs an information cache that works with an update-on-demand policy. Due to the multitude of databases and information sources, the most appropriate sources have to be selected for each query to ensure optimal and relevant data retrieval. To deal with this issue that the most suitable information sources have to be selected from a set of different distributed information sources that can provide the information needed, ActOn adds an information source selection step to the ontology-based information integration. Thereby, the most suitable information source database will be selected for a user query.

## *16.2.3 CDN Architecture*

Figure 16.1 shows three-tier view of CDN architecture. At the core of the middleware (in blue) lies the CDN ActOn Information Manager that represents the Cancer Data Content and its associated information extraction tools. The CDN ActOn contains the semantic metadata and knowledge management tools that enable modeling and analysing its life cycle and that support reasoning about the content. CDN also contains the workflow tools (Workflow engine) that enable the statistical analysis of the content enabling it to self-organize when linking with other content. Finally, the CDN also includes semantic-aware and peer-to-peer based networking functionality that enables the content to discover other content and communicate with it.

### 16.2.3.1 System Components

We use a bottom-up description of the components shown in Fig. 16.1.

**CDN Semantic Model**. The bottom layer represents existing and traditional data sources that will be used within CDN. Digital content elements on the sources will be identified and extracted and represented as Knowledge Cells (KC) that represent the core of an Data Content object and that represent nodes in abstract CDNs. Data sources can be accessed through the middleware and be offered to the application platform.

**ActOn Information Manager**. CDN middleware employs ActOn [1, 2], a semantic information integration system, to connect the data sources to the CDN system in order to: (a) deploy the Data Content and place it inside CDN which contain extra information about the content that make it both self-aware and context-aware together, and (b) to link the Data Content in multiple Data Content

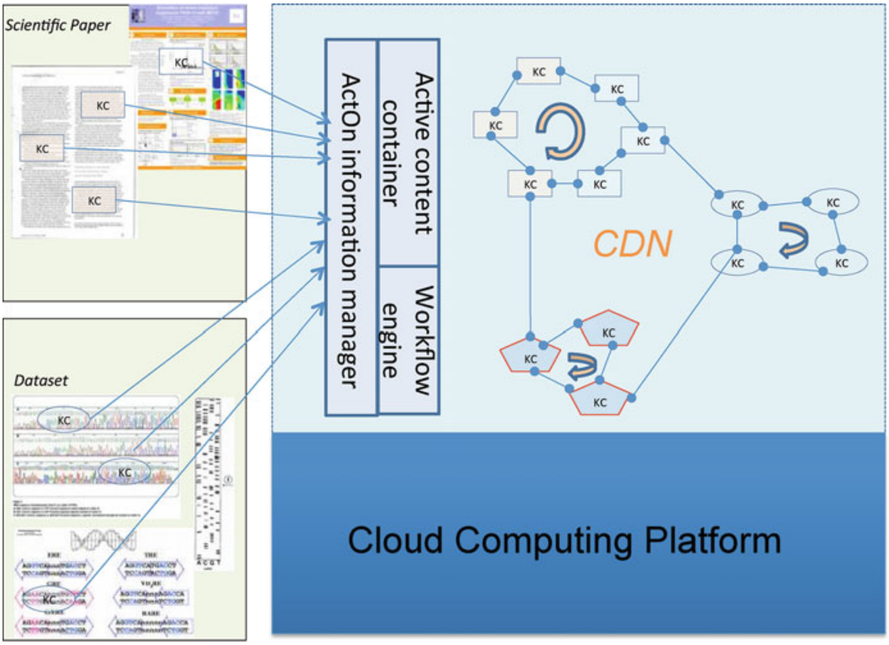

**Fig. 16.1** Overview of the CDN architecture

Networks. During system operation, the links (the edges in the network graph shown in Fig. 16.1) between Data Content entities (the nodes in the network graph in Fig. 16.1 can be re-organized based on statistical analysis, user preferences or other types of runtime information.

**Workflow Engine**. The Workflow engine enables data document access over preprocessing, tokenization, parsing, named entity recognition to the final consumer. It implements specialized workflows that support the different types of users of the system (data publishers, curators and end users) in combining data retrieval, integration, semantic annotation and deployment of Data Content within data analysis tasks in end user applications. The starting point for implementing the workflow engine is to employ Tarverna [3] workflow system (authoring tools and execution engines) for the integration and analysis of a wide variety cancer data (including genomic, proteomic data sets, as well as free text publications).

**Semantic Web User Interface**. At the top-level, a user interface includes functions that can be used to connect to the CDN middleware in order to: (a) Retrieve content from the distributed data sources (shown as different databases in the bottom of Fig. 16.1) in order to create the relevant Data Content. When Data Content is created, it is passed to the CDN middleware in order to be deployed in CDNs. (b) Interact with the user, i.e. issue user-queries and get back results which CDN can present to the users in an advanced way, showing the Knowledge Cells (KC) inside the retrieved content. This way, the user can use the KC in order to issue/refine further queries or even browse based on a given KC in order to find similar KCs and iteratively refine his/her queries within CDN in order to get satisfactory results.

## 16.3 Related Work

In the life sciences area there are several systems available that add semantic annotations, primarily these are done through Medline or similar literature databases. Some examples include iHop2, WhatIzIt3 and EBIMed4, and BioAlma5 [4–6]. Entities (such as gene names, protein names, drug names) are recognized, and links are added, however a disadvantage is that the recognition is not **active**, it is done once, off-line, and is not active in the CDN sense. Since the semantic framework to recognize entity identity across different services is currently missing, these services all point to a small subset of the data that is available for these entities, i.e. these systems are like isolated silos, compared to CDN's model of self-organizing, distributed structure. Adding semantic markups to more structured data is a relatively new area that has not been systematically explored in the biosciences. Such a system would take as input structured texts, such as protein sequence files, or entire databases, such as the EMBL, PDB, etc., then would add database cross-reference information based, in a way similar to SRS or MRS, then also add semantic annotations, similar to iHop [6–8].

## 16.4 Implementation

We prototyped the CDN system (shown in Fig. 16.2) using Java Spring Framework and RDF Jena API. Spring is a software toolkit that can be used to program web-based application and data management system. Jena is a Java API that can be used to create and manipulate RDF models. By using Spring framework, we are enable to code the system in Java following the WSRF specification. We use Jena OWL toolkit for creating, manipulating and querying the semantic metadata of Data Content.

Given the volume of the cancer data is large, we use CELAR cloud platform, an elastic cloud computing platform [9], to process and build the CDN networks. The EU CELAR cloud can deliver a fully automated and highly customizable system for elastic provisioning of resources within cloud computing infrastructures. It therefore can provide large-scale computation resources required by CDN. In addition, the CELAR platform can also provision particular types of computing resources

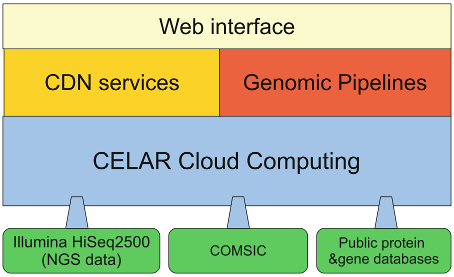

**Fig. 16.2** The CDN implementation

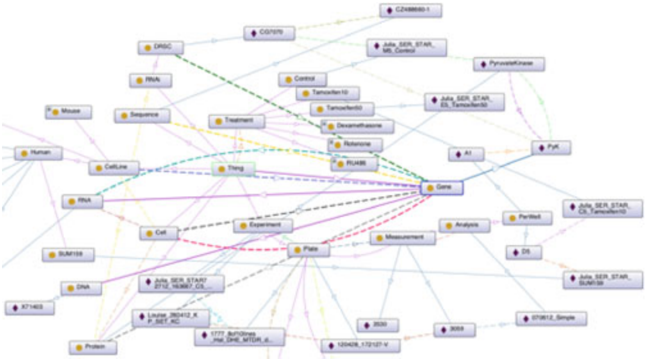

Fig. 16.3 An example of CDN network

required by CDN dynamically, such as windows system with large memory or large amount CPU resources of linux systems, etc. Currently our prototype implementation is mainly for creating and managing gene mutation data and the NGS variation detection data. The initial results are shown in Fig. 16.3.

## Conclusion

In this paper we have presented CDN an active content management system for personalized medicine research. CDN is based on a semantic content network approach which overcomes some of the limitations of current other content management approaches when dealing with dynamic, distributed and redundant bio-data sources.

Our main contribution over the state of the art in content management systems is that we have proposed Cancer Data Content architecture supporting deployment of Cancer Data Content, defining Cancer Data Content containers and the networking capabilities that allow remote interactions between Data Content entities. We also prototyped CDN as a cloud-based, networking middleware for Cancer Data Content discovery and communication.

The initial results show that CDN can facilitate both the cataloguing of samples collected during routine research and manage datasets generated by numerous multi-step experiments carried out from a single sample. For example, the CDN can provide a platform whereby all tissue samples, experimental step samples, datasets and analysis can be compiled and linked allowing ease of access to every stage in an open manner in order to streamline research and increase productivity.

**Acknowledgements** This work is supported by EU FP7 CELAR project.

## References

1. Xing W, Corcho O, Goble C, Dikaiakos MD (2010) An ActOn-based semantic information service for grids. J Future Gen Comput Syst 26(3):324–336. doi:10.1016/j.future.2009.10.003
2. Xing W, Corcho O, Goble C, Dikaiakos M (2007) Active ontology: an information integration approach for highly dynamic information sources. In: European semantic Web conference 2007 (ESWC-2007), Innsbruck, June 2007 (Poster)
3. Oinn T, Addis M, Ferris J, Marvin D, Senger M, Greenwood M, Carver T, Glover K, Pocock MR, Wipat A et al Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics 20(17):3045–3054
4. Rebholz-Schuhmann D, Kirsch H, Gaudan S, Arregui M, Nenadic G (2006) Annotation and disambiguation of semantic types in biomedical text: a cascaded approach to named entity recognition. In: Proceedings of the EACL workshop on multi-dimensional markup in NLP, Trente
5. Bioalma (2009). http://www.bioalma.com
6. Fernandez JM, Hoffmann R, Valencia A (2010) ihop web services family. In: JBI, pp 102–107
7. Sequence Retrieval System (2010). http://srs.ebi.ac.uk
8. Hekkelman ML, Vriend G (2005) Mrs: a fast and compact retrieval system for biological data. Nucleic Acids Res 33(Web-Server-Issue):766–769
9. EU CELAR Project (2013–2015). http://www.celarcloud.eu/