An extensible repository supporting clinical research in neuroscience

Corradi Luca\textsuperscript{1,2}, Porro Ivan\textsuperscript{1,2}, Schenone Andrea\textsuperscript{\ast1}, Momeni Parastoo\textsuperscript{3}, Ferrari Raffaele\textsuperscript{3}, Nobili Flavio\textsuperscript{4}, Ferrara Michela\textsuperscript{4}, Fato Marco\textsuperscript{1}

\textsuperscript{1} University of Genoa, Dept. of Communication Computer and System Sciences, Genoa, IT
\textsuperscript{2} Nextage srl, Genoa, IT
\textsuperscript{3} Texas Tech University, Lubbock, TX, USA
\textsuperscript{4} Clinical Neurophysiology, Dept. Of Neuroscience, Ophthalmology and Genetics, University Hospital S. Martino, Genoa, IT

Email: Corradi Luca - luca.corradi@unige.it; Porro Ivan - ivan.porro@unige.it; Schenone Andrea\textsuperscript{\ast} - andrea.schenone@unige.it; Momeni Parastoo - parastoo.momeni@ttuhsc.edu; Ferrari Raffaele - raffaele.ferrari@ttuhsc.edu; Nobili Flavio - flaviomariano.nobili@hsanmartino.it; Ferrara Michela - Michela3ferrara@hotmail.it; Fato Marco - marco.fato@unige.it;

\textsuperscript{\ast}Corresponding author

Abstract

Background: This work presents the architecture of an extensible software platform supporting a pilot study initiated by a multidisciplinary group of neuroscience collaborators at three different institutions. We have developed a novel approach in order to address technical challenges associated with: (i) developing and maintaining a database to harbor large clinical and genetic data implementing an extensible data structure able to include additional data types without the need for reconfiguration of the database (ii) merging information stemming from stored data and metadata and giving more comprehensive outline of all dataset; (iii) managing multimodal clinical files describing patient-centered diagnostic processes with complex architectures; (iv) utilizing statistical algorithms by exploiting distributed computational resources to analyze the stored data more efficiently.

Methods: The paper describes a comprehensive architecture for databases hosting clinical and research data. The architecture has been designed to meet two major requirements, namely, flexibility and extensibility. To fulfill those requirements, starting from existing standards, ontologies and reference data models available in neuroscience and bioinformatics, we developed a process-event data structure deriving from a highly flexible data model. Also, we designed and developed an innovative data storage system, in order to provide extensibility.
both, in terms of data input and data storage scalability, supported by a distributed DataGrid. The driving principle of having no preconfigured data type has been satisfied. It is up to users to configure the repository for the given experiment or data acquisition program, thus making it potentially suitable for customized applications.

**Results:** The goals of this project have been achieved through the design of a novel web-based extensible system that (i) supports researchers in the medical field to carry out data analysis using the integrated bioinformatics services and tools; (ii) handles multimodal/multiscale data and metadata, enabling the injection of several different data types according to structured schemas; (iii) is highly extensible, in order to address different requirements deriving from a large variety of applications simply through a user runtime configuration.

---

**Background**

Neuroscience is the study of all aspects concerning the anatomy and function of the nervous system, both in health and in disease. Neuroscientists use different tools ranging from classic clinical and neuropsychological evaluation and neuroimaging, to molecular genetics and cerebrospinal fluid (CSF) and blood biomarkers’ screening to better understand (i) the development and function of the nervous system in normal conditions and (ii) the dynamics of degeneration, such as in the case of neurodegenerative disorders. Integrating clinical, imaging and molecular data in robust, extensible and distributed databases represents a substantial challenge to modern neuroscience. Large datasets of different type and source (clinical, neuropsychological testing, imaging, molecular genetics, blood biomarker screening, etc.), as well as associated metadata, standard formats and similar data need to be collected, stored, monitored and analyzed. As a consequence, effective hardware and software environments must be provided to support data integration and analysis that ought to establish correlations among the totality of collected data.

According to the International Neuroinformatics Coordination Facility (INCF), neuroinformatics is the research field that encompasses the organization of neuroscience data and the application in neuroscience of computational models and analytical tools [1]. In this regard, our approach is intended to provide a flexible and extensible software environment aimed at supporting clinicians and researchers in managing large multimodal neuroinformatics projects.
The data challenge

Integrating several disciplines and dealing with different information and data represents a big challenge in terms of data management. In fact, data range from simple textual description to time series (Electroencefalography (EEG)), anatomical data (Computer Tomography (CT), Magnetic Resonance Imaging (MRI)) and functional data (Functional MRI (fMRI), Positron Emission Tomography (PET)/Single Photon Emission Computer Tomography (SPECT)). Data size and type can vary from few Kilobytes of a textual annotation on a patient, or a psychological questionnaire, to the hundreds of Megabytes of modern MRI/CT study or all day long multichannel EEG recordings [2].

Molecular genetics and the investigation of the whole genome by means of techniques such as whole genome sequencing, whole exome sequencing or genome wide association studies generate a large amount of digitized data [3] that need to be managed through ever developing bioinformatics tools, at the price of more and more data, metadata, and storage space to handle. As a response to these challenges, software environments and tools must be scalable in data modalities as well as in storage and computational resources.

The information challenge

Data management in neuroinformatics needs to provide not only adequate infrastructures to store, search and retrieve large data sets, but also tools and standards for defining their structure, exploring their contents and understanding their meaning. In other words, scientists need to focus on aspects such as accessibility, integration and annotation of data. It is even more difficult to provide extensible software environments able to effectively target the rapidly changing data requirements and structures of research experiments. In these cases, software environments should be flexible enough to allow the definition of new types of data and metadata, to support different associations between data and metadata, to modify data models and structures, to redefine the search paths and to add distributed sources of data.

State of the art

Several data models have been defined to describe structured data for storage in data management systems for biomedical applications. The main aim of such data models is to provide the definition and format of data for supporting the development of information systems.

XCEDE (XML-based Clinical and Experimental Data Exchange) [4] is a widespread data model providing an extensive metadata hierarchy for describing and documenting research and clinical studies. The
subdivision of experimental data at various granularity levels is granted by a taxonomy of hierarchical data
types. At each hierarchical level, elements contain level-specific information, whose schema may be used to
store metadata specific for the experiment or related to the data modality. The linking rules between the
levels are flexible enough to allow the omission of other levels if the user finds them unnecessary. Due to its
powerful general schema and to the inherent flexibility of its approach, the XCEDE model, describing a
complete metadata hierarchy for clinical experiments, has been taken as a starting point for building the
data structure of our repository. Some elements at the base of the XCEDE model (projects, subjects,
visits, studies, episodes and acquisitions) have been extracted and further developed in our work together
with some data and metadata hierarchies used by the BIRN (Biomedical Informatics Research Network)
initiative [5] [6].
The eXtensible Neuroimaging Archive Toolkit (XNAT) [7] is a powerful software platform designed to
facilitate the management and maintenance of neuroimaging data and related information developed by
the Neuroinformatics Research Group in collaboration with the National Institute of Health (NIH). As
XNAT implements a workflow to support the quality, integrity, and security of data from acquisition, and
storage of analysis and public sharing, it was considered as a reference platform in our work. However,
after a comprehensive testing and analysis of the code this approach has been proven to be impractical in
our application. The major obstacles were related to the manipulation and extension of the code, the low
flexibility in the data management and the difficulty to interface with distributed computing resources and
storage facilities.
Also the Functional Genomics Experiment (FuGE) object model aims at providing a framework for the
development of standards in life science. As such, FuGE provides a solid foundation for other
technology-specific, life-science standards and data formats and is currently being used to develop formats
for microarrays, proteomics, metabolomics and various other technologies [8]. References to FuGE have
been considered during our work.
From another point of view, ontologies try to define, through a common vocabulary, a structured
representation of the knowledge that can be used by either humans or automated software agents on a
particular domain [9]. Within this work two ontologies belonging to the Open Biomedical Ontologies
(OBO) [10] [11], providing a list of other useful ontologies, have been taken into account. They are:

. The Ontology for Biomedical Investigation (OBI, formerly called Fugo (Functional Genomics
  Ontology) [12] [13]
To address issues about distributed resources regarding both storage and computation, the Grid approach has been a successful way to manage some specific challenges about complex experiments in the biomedical field [16]. These challenges include the management of storage resources using different access protocols, the authentication and authorization across systems using different identity management systems, uniform management policies across institutions having differing access requirements and the need of a wide-area-network access.

iRODS, the Integrated Rule-Oriented Data System, is a data Grid middleware system developed by the Data Intensive Cyber Environments research group. The iRODS system is a generic middleware that can be configured to implement any desired data management application based on a set of major components including a data Grid Architecture based on a client/server model controlling interactions with distributed storage and computational resources, a Metadata Catalog maintained in a database system for managing the attributes of data and state information generated by remote operations and a Rule Engine enforcing and executing adaptive Rules. iRods is distributed with an open source BSD license and has been chosen as the reference Grid middleware for our software environment.

**Methods**

The need of extensibility has been considered from two different points of view. The first one is bound to the possibility to easily customize and extend the experimental procedures in order to track each step of acquisition or analysis. This is achieved through a process-event model, a multipurpose taxonomic schema composed by two generic main objects: events and processes. The second one is related to the improvement of data flexibility. This aspect has been taken into account through the development of a methodology for the dynamic creation and use of data types and related metadata, based on the definition of a "metadata model". This issue is critical in order not to constraint the repository to a set of predefined data but to make it easily extensible and applicable to different contexts, thus making data immediately usable and integrated. Finally, data integration aspects have been addressed by efficiently storing data and metadata and providing the repository application with an efficient dynamic interface designed to enable the user to both easily query the data depending on defined datatypes and view all the data of every patient in an integrated and simple way.
The process-event structure

An event is defined as either any “atomic” operation that can be performed on patients, or any processing of data, or any other action related to the administration and management of the repository. If needed, it can contain correlations between data, metadata and, in addition, algorithms. Each event is associated with a process. The process is defined as a group of sequential events and/or sub-processes related to an activity, allowing the creation of a sort of hierarchical structure. Custom process-event types and their relationships can be defined, thus describing the taxonomy better fitting the needs of the application. An example of the process-events structure concerning a possible clinical scenario is shown in Figure 1. It is worth noting the existing relationships between processes (blue boxes) and events (yellow boxes) and the association of data and metadata to each specific event. A pre-surgical analysis sequence is considered. This is divided into different phases consisting, each, of different steps, ranging from the acquisition of data to their analysis. According to the described structure, the Pre-surgical Process (P) can be considered as a top level process composed by different sequential subprocesses: (SP1) Data Acquisition, (SP2) Image Post-Processing, (SP3) Trajectories study and (SP4) Surgery Area Estimation. Each of these sub-processes represents a specific part of the main process and it is composed by a number of events each of them connected to related data and metadata. It is also worth noting that the whole process can be easily modified to fit either changes in the analysis sequence or different requirements for another case study, just by changing or adding new events, creating new processes composed of different events or combining existing processes and events. The defined process-event taxonomy can be used to store the information about each step in the process and the related data and metadata, thus allowing the definition of a detailed time line of performed operations. This can also improve the repeatability of experiments by providing both a detailed
recording of the analysis process and a complete description of relationships between data and actions. The XCEDE model has been tailored to fit this process/events model in a very simple and efficient way. As shown in Figure 1 the visit and study elements have been associated to process entities with a father-son relationship while episodes and acquisitions have been collapsed into events.

The data model

The analysis of the state of the art pointed out that several data models/repositories have the ability to add new data types. However, since their databases are based on the already defined data, they require the rebuilding of the database each time a new type of data is added. In this respect, we provide a new methodology improving the overall extensibility and customization and, based on the definition of a meta-“data model”, enabling the user to build his own data types independently from the application context. The presented approach, however, does not preclude the use of existing and standard data models. Indeed, through the development of suitable wrappers it is possible to convert data into a standard compatible format as well as to create types of data from an existing data model. Furthermore, the presented model is general enough to be also extended in order to describe other entities like experiment descriptions.

A "data type" is identified as a minimum set of information describing a data instance (i.e., the set of records associated to a clinical study, the parameters for a particular biomedical image) that may or may not be associated with physical files. As an example, clinical data can be defined as datatype but are not associated to a file, unlike MRI data.

Each data type is described by an XML metadata schema associated to XSD and XSL files file to define, respectively, its structure and display. The XSD and the XSL adopted for transformations are the same for all XML files. From XML files, DHTML web forms are built, using XSL transformations. The XML can be stored in MySQL or in other SQL databases. In this case, an XML file URL can result not in a physical file but in a query to the DB (this is transparent to the component requiring the XML). This can be considered as a caching mechanism. The XML representation of a data type metadata is divided into two main sections: a header, containing general information about the schema, and the metadata description, representing the detailed description of the information.

The metadata description is composed by one or more groups of information each identified by attributes, loops and their combinations.

An attribute defines a single metadata and is characterized by different parameters and subelements. The
formers describe the information related to the typology of the attribute (type, whether it is required, etc.). Subelements describe what the attribute represents; they include name, value or possible values and references to existing ontology definitions.

**Overall system architecture**

The main components of the Repository are (Figure 2):

- the Repository portal: it provides a web interface and allows users to access and manage database requests and is hosted by a Linux/Unix server environment.;

- the Database: it hosts all information about projects, subjects, metadata, etc.;

- the Grid Storage: it contains all data files;

Two important aspects are related to authentication and software as a service issues. As regards authentication, users and system administrators authenticate to the system using an existing LDAP or database account available on the server infrastructure. The access is via web browser without any client installation and in a secure way through the HTTPS (secure HTTP) protocol. When a user needs to access
the repository resources he has to authenticate himself through the web portal interface using his username and password. Each user is associated to Access Control Lists in order to guarantee security and auditing. System administrators are able to define different groups of users associated with different access permission to different pages and functions of the repository. In this way users can see only a subset of pages and perform a limited number of different actions (depending on their role in the project). The Web based SaaS (Software as a Service) approach has been preferred because the backend hardware can be scaled up enough to satisfy the user needs without requiring users to implement their own infrastructure. Moreover, this allows, at any moment, to migrate from a standard hardware infrastructure to a service based one, e.g. using the Grid or other kind of on demand IT services.

**The repository portal**

The repository portal is designed to make the storage and the navigation of data and information easy, through a simple and transparent web interface. It is a Java 2 Enterprise Edition (J2EE) web application based on several existing open source tools for the development of web applications. The basis of the portal consists in a framework that relies on a Apache Tomcat web application container [17]. It incorporates a database interface layer built through iBATIS, a persistence framework which automates the mapping between SQL databases and objects in Java [18]. To provide users with highly interactive interfaces, some components are designed using the AJAX (Asynchronous Javascript and XML) programming technique. Messages are exchanged in XML or JSON (JavaScript Object Notation) [19] format wherever possible. Wherever possible, XSL transformations (to transform XML data into human readable HTML pages) are performed [20].

This component represents the main access point to all the functionalities available through the overall integration platform, and exposes both user and administrator interfaces. Administrators are able to control users’ access by creating groups and their association with pages and functions, define processes (visits and studies), events and all their relationships, define new data types and related metadata, associate them with the related events and manage available ontologies. Normal users, according to their assigned permissions, are able to insert new data, retrieve patients’ information and view all the related data, download stored data, explore visits, studies and their interconnection and all the related events, data and metadata to have a global picture.

As an additional feature, in order to make the insertion of metadata easier, an automated approach is available for some predefined datatypes like MRI, fMRI, PET and SPECT. Using libraries like Java
dcm4che [21], the portal can automatically extract metadata contained within uploaded data files and incorporate them correctly in the database, creating associated events depending on image modality. Such as automatic procedure permits to avoid human errors and provides a further file type checking before the uploading.

If needed, a visualization tool can be made available within the portal interface in order to allow users to interact with neuroimages through a remote visualization service. This is made possible by using a client-side application that uses the VNC protocol to connect to a sharable work session that is running server-side, with a significant speed up of diagnostic processes.

The database

The Repository is based on a MySQL database. The database design has been a crucial part of the repository development. In fact this component is fundamental in order to make the repository highly flexible and easily extensible. The core of the database is formed by the two previously described entities: processes and events and their relationship to data and metadata. The information inside the data table represents the data inserted in the repository. These data can be associated with one or more files, thus keeping the association with one or more file entities accordingly to their datatype. The File table contains the URI of all the stored files. The repository can be configured to store the metadata totally or partially within the database. In this case the metadata are stored as XML descriptions inside the data table, to display the data in a rapid and dynamic way using XSLT Transformations and as records of specific metadata tables, to perform complex queries in an easier way.

The Grid middleware

A crucial aspect of the the repository design is related to the choice of the Grid middleware used to build the underlying infrastructure. We decided to use iRODS as the basis of the storage architecture of the repository principally because it directly handles not only data but also the related metadata, permits to use heterogeneous storage resources and allows the creation of microservices and rules to easily perform operations on the stored data and metadata.
Results
The scenario

Our platform is being tested within a research project currently carried out through an International collaborative effort between the Laboratory of Molecular Neurogenetics at the Texas Tech University Health Sciences Center (TTUHSC) and two Departments at the University of Genoa, Italy: the Department of Neuroscience, Ophthalmology and Genetics (DINOG) - Section of Clinical Neurophysiology and the Department of Communication, Computer and System Sciences (DIST).

The aim of the study is to develop criteria for the early diagnosis of AD and algorithms to predict the progression of the disease, by combining neuropsychological tests, imaging methods, genetic tests and biomarker screening. The steps identified within the collaboration in order to achieve these goals are as follows:

- Identify and enroll subjects: DINOG recruits patients with the diagnosis of single- or multi-domain aMCI (amnestic Mild Cognitive Impairment), Alzheimer’s disease and normal controls of a defined geographical area (Genoa, northern Italy). The clinical evaluation of patients is based on an initial interview, a complete general medical examination and a battery of neuropsychological tests. Further, MRI and fluorine-18 (F-18) fluorodeoxyglucose-PET scans are performed.

- Perform genetic and biomarkers screening: TTUHSC investigates the genetic makeup and the genetic markers of the study subjects to identify the intrinsic factors involved in the pathogenesis of dementia.

- Collect data: DIST collects and stores, over time, the results of clinical investigations and brain scans, together with the genetic and biomarkers screening in the multimodal and multiscale repository.

- Analyze multimodal data: Together with the physicians and the geneticists, specific algorithms will be developed to analyze and combine the acquired information. The generated algorithms will be valuable tools for the early detection of the disease and the prediction of disease progression. Figure 3 summarize the overall collaboration, roles and contribution of each partner.

From the architectural point of view, DIST provides the repository web application to all partners exploiting computational and storage resources through, respectively, an hybrid cluster and iRODS resources. Concerning storage, it is worth noting how data are distributed through different centers. In details, a main iRODS server is installed within the DIST infrastructure and further storage resources are located at Texas Tech (Figure 3). A Data Grid paradigm allows to keep the access and storage of the data unified, secure and transparent in resources geographically distributed among different centers.
Repository administrators may create new processes and associate them with sub-processes and events. Administrators can either modify already defined processes, by selecting them from a list, or creating new ones in an easy way, by means of the interface shown in Figure 4. Available categories for processes are: Neuro/Clinical data acquisitions, Samples management, Molecular genetics and biomarkers data acquisitions. Once a data type has been built, related XML file and database structures are automatically created and stored as described before.

After neurophysiologists have identified subjects eligible for the study, they have to insert the related personal data into the system together with clinical data, neuropsychological tests and neuro imaging data. The access to patient information is restricted because of privacy policies and only authorized users can insert, view or modify data. Clinical data and neuropsychological tests are simple data without any file association. Finally these users provide the blood samples drawn from the subjects. These samples are sent to the Texas Tech group to generate genetic data. In order to render samples management easy an ad hoc interface has been created to add samples, associate them with the corresponding patients and store related information.

At first, geneticists must configure the genetic setup by inserting information about the genes to be screened, their exons and their relevant known variants. This configuration can be performed through the integration with online resources such as NCBI (National Center for Biotechnology Information) [22] and Molgen (Alzheimer Disease and Frontotemporal Dementia Mutation Database) [23] in order to retrieve...
Figure 4: Interface to define new data types

and/or compare information. The user can directly view/access both the NCBI and the Molgen sites, the first pointing to the gene page and the second pointing to the gene variants section, in order to retrieve and store the correct and complete information within the repository.

A second task is related to the insertion of new data obtained by genetic screening on the patients’ samples. Finally, geneticists have to complete the samples management process by inserting information about the sample receipt and their storage. Concerning this last aspect, the repository is able to interact with an external application devoted to the samples storage and tracking within the fridges of the Texas Tech group. By means of such an integration, authorized users are able to manage samples and their storage information through a unique application.

Users can view more or less information according to their permissions: for example geneticists users are not be able to see patients personal data but only their clinical data. It is possible to combine common parameters like diagnosis, sex and birth date with data information in order to compose specific queries. It is in fact possible to add many different conditions in a dynamic way. As such, by choosing any of the defined data types, the corresponding ”field” parameter will be automatically filled with its specific defined attributes. Once a field has been selected, it is sufficient to insert the desired ”value” parameter to complete the condition. This method makes it possible to query patients having data with one or more specific metadata attributes for any arbitrary combination of their values. It becomes easy to imagine how to build complex queries choosing and combining common values (diagnosis and sex) with complex conditions on specific data and metadata. For example a query could be built to retrieve all male Amnestic MCI patients that are smokers, having variants for MAPT gene and one or more MRI data. After the

13
Figure 5: Interface to view and browse data of the patients have been retrieved in an integrated way. The red box highlights the visual integration between genetic and clinical data allowing to have an immediate overview.

query has been performed and patients have been retrieved, it is possible to have a quick overview of all their data using an interface divided into three main sections. The first one is the same used to build the query as described above and can be used to refine the query. The second section contains the list of retrieved patients. For the selected patient in the list a third section shows a synthetic view of the related data. Each field can be further examined in details.

Discussion

This work has been focused on data integration issues through the creation of a multimodal multiscale repository able to interact with external data sources. As a final result, all the main objectives described previously in the paper have been achieved. A web-based repository has been developed to collect multimodal and multiscale data and to access patient studies. Through the definition of a meta-“data model”, aspects related to data and metadata management have been handled. Particular attention was, in fact, given to issues related to the definition and development of a methodology to create and use dynamic data types and their metadata through appropriate XML schemas. This issue has become critical not to limit the repository to a set of predefined data but to make it easily extensible and applicable to different contexts and to make data being readily usable and integrated. The adopted methodology is also able to automatically generate interfaces to enter data and metadata by using XSL transformations, and to perform complex queries composed on the basis of stored data and metadata. Extensibility issues have been also managed through the definition and
implementation of the process-event model, a multipurpose taxonomic schema designed to easily customize and extend the experimental procedures in order to track each step of acquisition or analysis. Such a model fits into the neuroscience context due to the adoption of concepts defined within the XCEDE model (project, visits, studies, episodes, acquisition). A Grid approach has been considered and implemented in order to manage distributed, heterogeneous data and information, improve security policies and facilitate collaborative work. Furthermore the Grid paradigm enables a safe access to data stored on distributed resources by using an X.509 based secure access. Finally, the developed architecture and platform have been deployed and customized in order to fit the identified use case.

Conclusions

As future developments of the presented environment, next steps will be about: (i) the identification of analysis algorithms and visualization tools; (ii) their implementation/integration within the provided infrastructure by creating services able to analyse and view data through remote computational and visualization resources. Afterwards, a possible future development is related to the improvement of the process-event model. In fact such methodology is highly flexible and its conception can be developed beyond a taxonomic schema to describe the procedures. The described structured information could also be used and act as a reference to compose workflows of analysis. The planned improvement is related to the possibility of associating services to events. This would enable users to create simple or complex pipelines involving stored data and defined processes and events, convert the workflow in a standard format and execute it using an integrated workflow manager. Concerning the presented use case, data and results coming from clinical investigations, brain scans, together with genetic screening are currently being collected and stored over time within the repository. A first set of fifteen patients with their clinical data and brain scans has been inserted so far through our platform waiting for results of related genetic data. Then, specific algorithms to analyze and combine the acquired information will have to be developed and could be exposed as services. The related processes and events should be defined according to analyses to be performed and new data results to be stored.

List of abbreviations used

Text for this section.
Competing interests

Text for this section.

Authors contributions

LC participated in the design of the system, carried out the development of the code and helped to draft the manuscript, IP participated in the design of the system and carried out the development of the code, AS participated in the design of the system and helped to draft the manuscript, PM participated in the definition of genetics data types and structures and helped to draft the manuscript, RF participated in the definition of genetics data types and structures and helped to draft the manuscript, FN participated in the definition of neurophysiological data types and structures and helped to draft the manuscript, MFerrara participated in the definition of neurophysiological data types and structures and helped to draft the manuscript, MFato participated in the design of the system and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Part of this work has been funded by the European Commission FP7 Project Decide (Diagnostic Enhancement of Confidence by an International Distributed Environment).

References

1. International Neuroinformatics Coordination Facility[http://www.incf.org/].
4. XML-Based Clinical Experiment Data Exchange Schema (XCEDE)[http://www.xcede.org].
5. Biomedical Informatics Research Network (BIRN)[http://www.birncommunity.org/].


10. The Open Biomedical Ontologies[http://obo.sourceforge.net].


17. Apache Tomcat web application container[http://tomcat.apache.org/].

18. iBATIS[http://ibatis.apache.org/].

19. JSON (JavaScript Object Notation)[http://www.json.org/].

20. XSL Transformation[http://www.w3.org/TR/xslt].


23. Molgen (Alzheimer Disease and Frontotemporal Dementia Mutation Database)[http://www.molgen.ua.ac.be/Admutations].

**Figures**

**Figure 1**

XTENS process-events structure example

**Figure 5**

XTENS overall architecture

**Figure 8**

Partner roles and contribution summary

**Figure 9**

Interface to define new data types
Figure 12

Interface to view and browse data of the patients have been retrieved in an integrated way. The red box highlights the visual integration between genetic and clinical data allowing to have an immediate overview.
### New Data Type Creation

**Data information**

<table>
<thead>
<tr>
<th>Name</th>
<th>URI</th>
<th>File Upload</th>
<th>YES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>URI DATA</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Process-Event Association**

<table>
<thead>
<tr>
<th>Event Type</th>
<th>URI ACQUISITION</th>
<th>Process Type</th>
<th>NEUROCLINICAL DATA ACQUISITION</th>
</tr>
</thead>
</table>

**Metadata information**

<table>
<thead>
<tr>
<th>Version</th>
<th>Ontology</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>01 00:00</td>
<td>Ontology for Biomedical Investigations (OBI)</td>
<td></td>
</tr>
</tbody>
</table>

---

**Group 0**

**Name** GENERIC INFORMATION

**Attribute**

- **Type**: STRING
  - **Name**: INSTITUTION_NAME
  - **Value**: [Input field]

- **Type**: STRING
  - **Name**: MANUFACTURER
  - **Value**: [Input field]

- **Type**: FLOAT
  - **Name**: MAGNETIC_FIELD_STRENGTH
  - **Value**: [Input field]
**Clinical Details**

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>USE OF ALCOHOL</td>
<td>NO</td>
</tr>
<tr>
<td>IS SMOKER</td>
<td>YES</td>
</tr>
<tr>
<td>IS SMOKER DATE</td>
<td></td>
</tr>
<tr>
<td>EX SMOKER</td>
<td>YES</td>
</tr>
<tr>
<td>DIABETES</td>
<td>YES</td>
</tr>
<tr>
<td>HIGH BLOOD PRESSURE</td>
<td>YES</td>
</tr>
<tr>
<td>HIGH BLOOD PRESSURE DATE</td>
<td></td>
</tr>
<tr>
<td>HIGH COLESTEROl</td>
<td>NOT DECLARED</td>
</tr>
</tbody>
</table>

**Molecular Genetics**

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>MAFT</th>
<th>Known Variants</th>
<th>Found Variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACE Haplotype</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs16832753 - C&gt;G</td>
<td>H1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAFT Haplotype</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs16832753 - C&gt;G</td>
<td>H1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>APOE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs6759589</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs7432940</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Data Table**

<table>
<thead>
<tr>
<th>#</th>
<th>Code</th>
<th>Surname</th>
<th>Name</th>
<th>Sex</th>
<th>Birth date</th>
<th>Diagnosis</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CNTU57L42D099Y</td>
<td>COTTE</td>
<td>OLIVIA</td>
<td>F</td>
<td>02/07/1937</td>
<td>ALZHEIMER</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>COTLUCY91D14D099E</td>
<td>COTTO</td>
<td>LUCA</td>
<td>M</td>
<td>14/04/1981</td>
<td>CONTROL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>PGGMP28SH140D099F</td>
<td>POGGI</td>
<td>MAURIZIO</td>
<td>M</td>
<td>14/06/1985</td>
<td>NON-AMNESTIC MCI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>ROSSOR65D099R9F</td>
<td>ROSSI</td>
<td>GIORGIO</td>
<td>M</td>
<td>09/04/1989</td>
<td>CONTROL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>RSSMR181D14D099Q</td>
<td>ROSSI</td>
<td>MARIO</td>
<td>M</td>
<td>14/04/1981</td>
<td>AMNESTIC MCI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>RZZSNN7L1D099NL</td>
<td>ZUZEN</td>
<td>SIMONA</td>
<td>F</td>
<td>02/07/1977</td>
<td>NON-AMNESTIC MCI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>