© 2013 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

e-Enabling International Cancer Research: Lessons Being Learnt in the ENS@T-CANCER Project

Anthony Stell, Richard Sinnott
Melbourne eResearch Group
Department of Computing and Information Systems
University of Melbourne, Melbourne, Australia
anthony.stell@unimelb.edu.au

Abstract—**Breakthroughs in biomedicine are driven by research. More often than not, research takes place outside of a healthcare setting. However access to and use of clinical data for research purposes has many challenges that must be overcome, not least of which are the lack of standardized nomenclature and the heterogeneity of healthcare IT systems. For rare conditions, this challenge is particularly acute since the scarcity of data makes scientific breakthroughs increasingly difficult. Adrenal tumours represent one rare disease area where consolidation of clinical and biological information is urgently required. This paper describes the lessons being learnt in the development and rollout of an advanced security-oriented, virtual research environment (VRE) as part of the EU funded ENS@T-CANCER project. This system is currently used by 39 major cancer research centres across Europe and provides a unique resource for adrenal cancer research, underpinning an expanding portfolio of major international clinical trials and studies.**

Keywords-adrenal cancer, virtual research environment, distributed biobanking

I. INTRODUCTION

Access to and use of clinical data for research purposes has many challenges. Not least of these are the lack of standardized nomenclature and the heterogeneity of healthcare IT systems. For rare conditions, this challenge is particularly acute since the scarcity of data makes scientific breakthroughs increasingly difficult. Adrenal tumours represent one rare disease area where consolidation of clinical and biological information is urgently required.

To help tackle this, the European Network for the Study of Adrenal Tumors (ENS@T - www.ensat.org) was founded in 2002 through the merging of three existing but largely independent, adrenal tumor research networks in France, Germany and Italy, with research teams from the UK. The central aim of the ENS@T consortium was to improve the prediction and management of specific types of adrenal tumors. In particular the community focused on the tumour types: adrenocortical carcinomas (ACC), pheochromocytoma and paragangliomas (Pheo/PGL), non-aldosterone producing adrenocortical adenoma (NAPACA) and aldosterone-producing adenoma (APA) - all of which are relatively rare and have typically poor survival rates. Building on the ENS@T network and the community of adrenal tumour researchers, the ENS@T-CANCER project was funded as part of the EU Framework 7 initiative in 2011. At the heart of the ENS@T-CANCER project was the establishment and support of an advanced clinical platform – a security-oriented, web-based virtual research environment (VRE). It was intended that this VRE would support all aspects of the collaboration including collection, sharing and analysis of biomedical data (both physical and digital) associated with adrenal tumours, and the interactions of the clinical and research communities involved. Already, this VRE has galvanized a community of adrenal tumour research efforts across the EU - far beyond the original ENS@T-CANCER project members. It offers a rich range of functionality supporting the biomedical and clinical communities. At present the VRE is used by 39 major clinical cancer centres from across Europe; it offers a consolidated repository of adrenal cancer phenotypic data from over 4250 patients with over 27,000 clinical annotations (treatments, surgeries, follow up information etc); supports advanced biobanking facilities and is used across a range of major international clinical trials and studies. The focus of this paper is to describe how and why this system has grown with the community demand, focusing on the lessons learnt and the capabilities that it offers along with the underlying technological challenges that have had to be overcome.

The rest of the paper is structured as follows. Section II focuses on the clinical context of ENS@T-CANCER and the capabilities required. Section III describes the lessons learned from, and the solutions provided to, the challenges posed by growing data in a real clinical setting. Section IV outlines the specific requirements of two studies that are heavily integrated with the registry (PMT and EURINE-ACT), and sections V and VI describe the evolving implementation and security solutions incorporated into the VRE.

II. CLINICAL CONTEXT OF ENS@T-CANCER

The goal of ENS@T and ENS@T-CANCER was to better understand the molecular mechanisms associated with adrenal tumours, and in so doing improve the treatment and ultimately the management of patients with adrenal tumours. As each individual is genetically unique, tumour types are also biologically unique. The much-vaunted vision of personalized medicine/e-Health applies directly to the work of ENS@T-CANCER [1]. Different individuals have different tumour types that have different biological manifestations and respond

to different treatment regimes. In some individuals, cancers can metastasize rapidly; in others tumours can be benign for many years before metastasizing; some patients respond well to certain treatments, whilst others do not. Understanding this at the individual and cancer subtype level is the goal of ENS@T-CANCER.

Prior to ENS@T-CANCER, adrenal tumor information was located in isolated national (and/or individual hospital) repositories of highly valuable data invisible to the vast majority of clinicians and researchers. The need for coordinated data sharing was essential. Contemporaneously, with the establishment of ENS@T-CANCER a multitude of clinical trials were identified and scoped with the ENS@T-CANCER initiative in mind. The need to recruit patients to particular studies and manage the associated physical (e.g. bio-specimens) and digital information from these patients was essential. These studies included the multicentre international clinical trials ADIUVO, FIRSTMAPPP, EURINE-ACT and PMT [2]) – all of which are now up and running and integrated into the ENS@T-CANCER VRE. Furthermore, other clinical trials and studies are now starting to appear (e.g. CHIRACIC, long-term PHPGL, and AVIS [2]). Whilst funded in their own right, coordination of these trials through the ENS@T-CANCER VRE is essential to ensure cohesion across the adrenal tumour landscape. For instance, prospectively enrolled patients in a particular study may have information useful to the central adrenal research community, whilst any information held in the registry can provide retrospectively enrolled patient data for a given study. Furthermore, given the rarity of the conditions there is a real chance of the same patients being recruited to separate studies, thereby potentially jeopardising the associated studies.

As one example of these, the PMT Study was designed as a full four-phase clinical trial led by the Technische Universität Dresden in Germany, and partnered with centres specialising in the study of pheochromocytomas in Warsaw (Poland), Nijmegen (Netherlands), Munich and Würzburg (Germany). The trial was tasked with prospectively recruiting 2400 patients that have suspected pheochromocytomas and track their progress over several years. The exchange of information between the core ENS@T-CANCER VRE and the PMT study required dynamic connections to the central Pheo and NAPACA registries. It was intended that a multitude of specific measurements of biochemical indicators were to be captured in the study, leveraging as much of the data and harmonisation of information established in the ENS@T-CANCER systems. It was essential to uniquely identify (link) patients individually through the registries, across studies and importantly with the samples/bio-specimens that they may have. These are typically stored in pathology laboratories in local hospital settings, which can have their own labelling and laboratory information management systems (LIMS) in place.

Biomaterial sample exchange was identified as one of the key areas where the VRE was required to provide interfaces with the real physical world. Monitoring of sample availability, tracking its progress between collaborating centers, and protecting the identity of the patient at all times were key. To facilitate the progression of clinical research in the adrenal tumour arena, the use of biomaterials themselves are essential. A range of groups with a range of bioinformatics capabilities exists in ENS@T-CANCER, all of whom have their own complementary –omics approaches to analyzing physical samples. The development of the VRE was thus required to focus on biosample sharing as a major priority. This required that many logistical considerations were tackled since every hospital has different systems and arrangements for labeling and storing bio-specimens, and indeed collects different physical specimens (normal/cancer tissue, DNA, urine, 24 hour urine, etc): Technical solutions were thus required for harmonising site-specific storage within laboratory freezers; ensuring and standardizing the viability of the labels used, e.g. tolerance of extreme low temperatures.

It was also recognized that some studies would have greater need for sharing of biomaterials than others. The EURINE-ACT study is a good example of a study heavily dependent on the use and transfer of biomaterial information. This is a prospective study that investigates the prognostic and diagnostic value of urine steroid metabolomics in patients with ACC using approaches based on mass spectroscopy (GC/MS) analysis. The investigation builds on previous work by the team at the University of Birmingham who have shown that a simple urine test may provide information of higher specificity and comparable sensitivity for cancer subtype identification to that of imaging techniques currently used.

A. *Related Work*

From a wider computer science/informatics point of view, there are a number of other initiatives in the space of supporting clinical research. Drives to link health data ("e-Health") and the amassing of large, representative biomaterial of a certain population are core goals of efforts such as CaBIG [3], BioGrid [4] and the UK Biobank project [5]. Whilst these projects have a large amount of effort and momentum, it can be argued that their scale and their scope impinge on their overall utility since they are designed to service the needs of many communities. The ENS@T-CANCER systems have been designed to address the needs of a very target audience, who have, and still are, driving its development in a tight iterative feedback cycle. The systems are now flourishing with adoption and utilisation by a multitude of clinical/biomedical partners beyond the original ENS@T-CANCER project partners.

Furthermore the ENS@T-CANCER VRE systems are now being adopted by the Australian National eResearch Collaboration Tools and Resources (NeCTAR) funded Endocrine Genomics Virtual Laboratory (endoVL – www.endovl.org.au) project also based at the University of Melbourne [6]. The ENS@T-CANCER systems are currently being being repurposed to other clinical registries and studies, e.g. the Australian Diabetes Data Network (www.addn.org.au) and the NHMRC funded ENDIA project (www.endia.org.au) with specific focus on the challenges of rolling out such systems in/across Australia.

III. LEARNING FROM THE GROWING DATA CHALLENGES

A number of informational hurdles exist in the development and maintenance of such a complex environment. This section outlines these challenges and the lessons being learnt in tackling them.

A. Identification, linkage and security

Central to the information model used in the ENS@T-CANCER project is the notion of a unique patient index that encapsulates the patient identity within a centre in a particular country. This identity is critical to maintaining the consistency of the information held within the registry but it must also be decoupled from any local institutional identifiers (such as hospital number of patient name/address etc) according to the specifications laid out in the (approved) ethics application for the project. Such identity mapping fits with the requirements of the ISO standard (27001) and specifically the section on identity management [7].

The numbering itself is facilitated through use of a centralised – separate – database that maintains a running count for each centre registered in the VRE. In centres conducting studies that are related to the ENS@T-CANCER VRE – for instance, the PMT study (Pheo), and the ADIUVO study (ACC) – the identification mapping, and the callout to this separate numbering database, has had to be dealt with on a case-by-case basis. Initially a simple numbering scheme was adopted (an increasing integer), however with the volume of patients and the international coordination a new scheme was adopted comprising country code, centre code and unique patient identifier.

As the work progressed, it was also recognised that flexibility in the center specific information was required, e.g. due to staffing changes. Given this, the security credentials evolved from fixed username/password per individual, to role-based access, which accommodates clinical staff leaving their posts. Furthermore an alias contact was required to be maintained per center rather than tying the record ownership hierarchy to an individual project investigator. This is also reflected in the peer-to-peer network of the security policy (see section VI). In addition to clinical staff turnover, a more common scenario – one that turns out to be more common due to the specialist nature of the field – is that clinicians move between centers that can both be involved in the consortium. Therefore, the association of a clinical professional with an institution must have a one-to-many relationship, which is not immediately intuitive when integrating access privileges for a specific center.

In the PMT study, the electronic Case Report Forms (eCRFs) and associated numbering schemes are under the control of the developers responsible for the ENS@T-CANCER VRE with full access and control to all of the databases involved. Whereas in the ADIUVO clinical trial, a web service interface has been implemented between the registry and the Italian-based study eCRFs – maintained by a separate set of developers, who work in collaboration with the ENS@T-CANCER developers to establish this connection and numbering scheme consistency. As the ADIUVO study commenced just before the ENS@T-CANCER initiative, the ADIUVO study had its own numbering scheme already in place. The way chosen to address this disconnect of two unrelated schemes was to maintain an updateable mapping file that relates the generated patient IDs. Similar issues existed in the bulk upload of legacy data, where the mapping of meta-data information was required between different international centres, most notably in the French (Access database) and German (Excel spreadsheet) ACC records. Once the critical mass of participating nations is established it is much easier to ask new nations to adopt the existing, useful standard.

B. Sample Exchange

Corollary to the notion of central identification numbers has been the ability to transfer patient data between the research communities involved in the collaboration as accurately and seamlessly as possible, and with as small an impact on the clinicians/researchers as possible. The scheme to achieve this was first prototyped and rolled out in production between the ENS@T-CANCER systems and the PMT study. This utilised a back-end database sub-schema common to both the PMT and ENS@T-CANCER systems. Due to the differing requirements of the ENS@T-CANCER systems and the PMT study, a net effect – acknowledged as acceptable by the clinicians involved on either side of the transfer – was the "dilution" of information as it passes from one system to the other. For instance, detailed information on biochemical measurements of plasma and urine metanephrine and methoxytyramine levels are critical in the PMT study and are thus highly specified. In the ENS@T-CANCER registry the overall outcome of the information – for instance, what the levels indicate clinically – rather than the detailed measurements, is the key information.

Another feature of the ENS@T-CANCER VRE is the ability for patients to be transferred from the PMT study – a study investigating pheochromocytomas tumour subtypes – into the NAPACA section of the registry. This is enabled to allow unspecified adrenal masses – which may yet still present as pheochromocytomas – to be entered into both the study (before confirmation or exclusion of the pheochromocytomas) and the ENS@T-CANCER VRE (as the adrenal mass can be of interest for those focusing on the study of NAPACA-type tumours). These considerations are also present in the transfer function developed between the ADIUVO study and the ACC section of the registry, though again the implemented mode of web service differs from the direct database connections used for PMT.

Key to these transfer functions is the ability to crosscheck the sample availability and those transferred between centers using a manifest file that can be easily printed and accurately associated with biomaterial shipments. The PMT study explicitly allows a summary manifest to be printed to an Excel spreadsheet within the application.

C. *Biobanking Features*

A prominent role of the ENS@T-CANCER registry is to provide a centralised biobanking facility that aids the exchange of sample and tissue data across international boundaries. There are core capabilities that have been identified as critical to making this application effective, and these have been developed in consultation with the centres that possess, distribute, process and analyse such data. For instance, the laboratories at the Technische Universität Dresden and the University of Birmingham have been identified as "high-throughput" centres – where many samples are processed using high throughput –omics data processing, e.g. LS/MS mass spectrometers. Other lower throughput data processing centres are also present, such as INSERM (Institut National de la Santé et de la Recherche Médicale) in Paris, France, and the CNIO (Fundacion Centro Nacional de Investigaciones Oncologicas Carlos III) in Madrid, Spain.

To standardise the biomaterial exchange, a core set of parameters have been laid out in forms that belong to a patient record. Since many of these forms can be required per patient, clinicians are strongly encouraged to associate each form with a particular study ("ADIUVO", "PMT", etc). When that study association has been made, the required items are pre-populated into the form to validate the form data. For instance, for the EURINE-ACT study, specific levels of urine, plasma and serum are required, which are flagged in red text on biomaterial forms that have been indicated as belonging to the study. A further critical component to the system is the ability to print labels that will then be affixed to the tissue sample tubes before they are shipped between centres. The information provided includes a high-level description of the contents, the relevant identification numbers, the aliquot number (where one sample is divided into aliquots for logistical purposes), and a barcode to allow ease of scanning by high-throughput machines.

Providing these features and their convenience value has turned out to be critical in the encouragement of clinicians to use the ENS@T registry as a biobank support facility.

IV. Learning from the Studies – PMT and EURINE-ACT

The two studies that have made heaviest use of the distributed biobanking facility and shaped the ENS@T-CANCER capabilities to date are the PMT (Prospective Monoamine Tumor) study and the EURINE-ACT study.

The PMT study is a full four-phase clinical trial that tracks patients who exhibit clinical indications of suspected pheochromocytoma through any known signs, hypertension, or incidental or screening discoveries.

Patients are admitted to the study from the specialist clinics in Germany (Dresden, Munich, Wurzburg), Netherlands (Nijmegen) and Poland (Warsaw), tracked through the four phases of the trial – screening, clonidine tests, characterisation, follow-up – and beyond up to five years after study. The study required 2400 patients to be recruited and have an extensive set of eCRFs developed for each phase of the clinical trial.

A key requirement for this study was the ability to analyse patient biochemistry (e.g. metanephrine concentration) from all patient records at the specialist laboratory in Dresden, which therefore requires the transfer of biomaterial samples from the other centres. Additionally, the supporting clinical information must be associated with those biomaterial forms for the correct patients. The unique combination of the PMT eCRFs and its link to the ENS@T registry, allows this transfer of information to occur in a feasible timescale for the completion of the study. The full standard operating procedures (SOPs) for the study can be found at the PMT study homepage [8] and details of the technology behind the eCRFs is given in [6].

The required inputs to the EURINE-ACT study are shown in diagram format in figure 1. The flowchart describes the input of clinical information after the inclusion criteria have been met (adrenal mass > 1cm), and the required biomaterial samples after the study proceeds. 24-hour and morning spot urine, plasma and serum samples are required at 3-month intervals, so the demand for sample transfer tracked efficiently between international centres is high. Therefore, again there is a strong need for a distributed application to support such transfers with the associated clinical information.

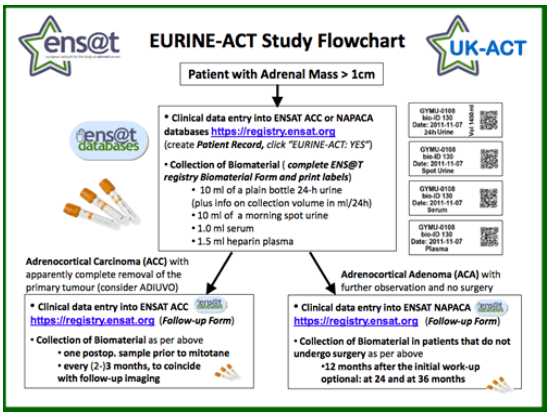

Figure 1: the required clinical and biomaterial inputs for the EURINE-ACT study

For both studies – and the others involved with the registry (ADIUVO, FIRSTMAPPP, etc) – the ability to harness all this information into a convenient format for data analysis (mainly Excel spreadsheets) and to trace the provenance of the data through accurate logging is one of the main advantages of combining this information with the use of a distributed biobank facility. The next section details how the registry is implemented and supports these requirements.

V. Lessons Learnt in Core Registry Features and the Evolving Implementation

The ENS@T-CANCER VRE has been built using a standard n-tier web application setup with a MySQL back-end database, business logic programmed in Java rendered to a JSP front-end – hosted in a Tomcat container – with various

libraries used to provide additional feature support (Apache POI, iTextPDF, Guava). Parameter rendering is achieved using a separate database listing the available parameters for the central forms and the subsidiary "one-to-many" forms implemented for treatments, biomaterials and follow-up. Standard tracking features are also currently implemented for both, outlined in section VI.

Having been involved in the establishment of a spectrum of clinical / biomedical studies [9] that have adopted other technological solutions the above systems are designed specifically to be simplistic. Whilst these solutions have shown that it is indeed possible to apply such technologies to deliver a prototype system, the ENS@T-CANCER VRE technological base has a primary focus in delivering the solution required by the clinical community, as opposed to the challenges in "making the technology work" focus of other endeavours.

The experiences of applying middleware frameworks (e.g. the Globus Toolkit) is that their development doesn't directly meet the community requirements, and the resultant overhead in tailoring the technological solution becomes unacceptably large. As a result, where frameworks are used (for instance, in the Google zxing project [10] for creating barcodes), they are focused to a very specific requirement (barcodes for the labels) and the rest of the application is not tied to using the same stack for other functions.

The high traffic use of the ENSAT registry also contributes as a primary driver of this approach. As feature and maintenance requests are being serviced daily, the generic features of a portal framework or middleware application are required to work in a streamlined and as seamless a manner as possible, and dependence on third-party support is required to be minimised (especially if these frameworks are developed by academic groups with limited funding support). These considerations have driven the choice of technology and approach described here.

A. *PMT Clinical Trial Interfaces*

The PMT eCRFs have also been developed and hosted by the Melbourne eResearch Group, and are described in more detail at [19]. This has allowed a parallel development of both study and ENS@T-CANCER core VRE structure that are cognisant of each other.

The main use-cases for patient record transfer are two-fold: for the ENS@T-CANCER to PMT transfer, it is a useful tool to incorporate patients eligible for the PMT study that are already in the ENS@T-CANCER Pheo database. For the reverse transfer, the primary driver is to note information about patients that have entered the PMT study, have turned out not to be presenting with pheochromocytoma and are eligible to be recorded in the NAPACA section of the registry. It is however also useful to transfer information about patients with confirmed pheochromocytoma tumors into the registry for future potential studies.

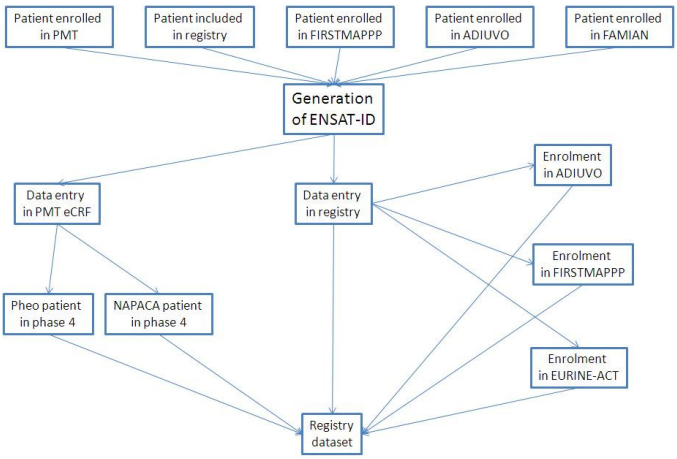

Figure 2: example of potential data-flow between the registry and contributing studies. Patients enrolled in PMT can go into the data-set via the NAPACA or Pheo sections. Patients from other studies can go into the registry (prospective recruitment) or the other way round (retrospective).

Figure 2 shows examples of the data-flow between studies and the central registry. Key to this is the generation of an ENSAT ID, which unambiguously records the clinical patient information without revealing identifying details (that remains within the center that owns the patient record).

B. *EURINE-ACT Clinical Trial Interfaces*

The EURINE-ACT study originally had local spreadsheets to act as the primary data-store, but now uses the central ENS@T-CANCER VRE for this purpose. The data requiring collection is simple, but there are some more complex features required for the study analysis, that are ideal for generalisation to other studies.

First is the heavy use of automatic email notifications to keep the principal investigators aware of eligible patient numbers. An email is sent when a patient is entered that is flagged as eligible for this study. An email is sent to the patient record owner when the metabolomics analysis is uploaded to the registry (see below), and email reminders are periodically sent to all clinicians involved in the study, detailing the outstanding biomaterial transfer requirements for their patients.

The second main requirement is a feature to upload a (PDF) file showing the steroid analysis of the patient's samples (figure 3). The ability to upload this information in this format is another generalisable feature for other studies, but with an attendant notion of fine-grained security in its use. This comes about due to the more sensitive research value of this analysis – though the collaboration is open, the ability to lock down such privileged information to only the interested parties in a convenient way is required.

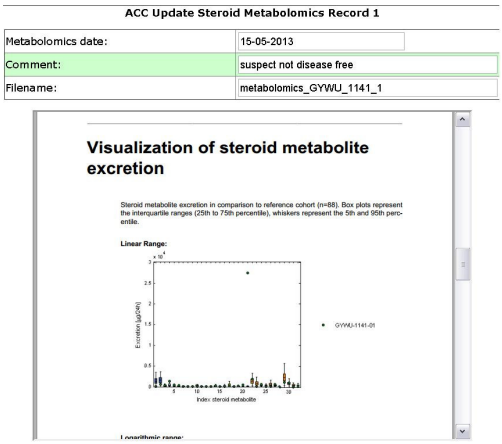

Figure 3: an embedded PDF file (using the iframe tag) in the registry showing the steroid profile of a patient record for the EURINE-ACT study.

C. *Biomaterial features*

The central information point in the registry, to enable the type of sample transfers described in previous sections, is the biomaterial form. Figure 4 shows the list of biomaterials types that are represented, which by consensus of the ENS@T working groups are believed to cover all of the potential samples that will be exchanged.

Figure 4: example biomaterial form in ACC (listings are uniform across all sections)

This form has various features implemented to support the unique requirements of biomaterial information storage.

1) *Freezers and aliquots*

To accurately track the physical specification of the biomaterial samples, key parameters to note are the aliquots that the samples are split into and their position in the local freezer. Sections in the form are programmed to open upon selection detailing the aliquot number, and the freezer, shelf, rack, box and position numbers therein (figure 5). If the position is already occupied then a validation flag notes this and prevents re-entry of that specific position. When the sample is used and the position becomes free, this is noted and archived in an archive history of the local freezer. In this way it is possible to corroborate the transfer of material between centres to a high degree of accuracy. The next step in development of this feature is to bring the transfer data together into one page detailing their in-transit progress between centres (similar to package-tracking systems used by courier services).

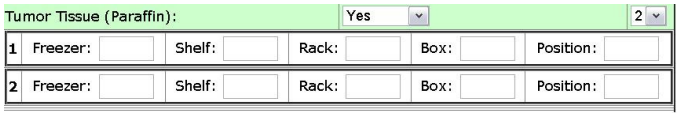

Figure 5: example of the aliquot and freezer specification in the biomaterial form. The two freezer lines correspond to the selection of two aliquots of the sample

2) *Labels and barcodes*

The production of labels is obviously of great significance when transferring human materials, along with the ability to provide information that is relevant to all of the participating centers. In this regard the sample specifications are stored permanently with the option for modification when it comes to the time of printing. The information captured includes the canonical identifier (including the center code), the biomaterial form number, the aliquot number, and the sample name. This is also captured in barcodes for readers that wish to capture the information electronically [10]. There are many barcode standards to choose from – QR codes were initially tried but rejected due to the two-dimensional nature being unreadable on the curved surface of a typical sample tube. As a result the project has now adopted Interleaved 2 of 5 [11] as the agreed standard (a label example is shown in figure 6).

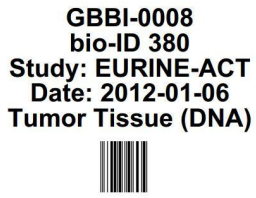

Figure 6: an example biomaterial label output, showing barcode, form and study information (in this EURINE-ACT)

3) *Study Association*

Combining these features together, one of the primary convenience features of using these biomaterial forms is the ability to select a particular study and the required features automatically populating. For instance, if “EURINE-ACT” is selected for the associated study, the parameters required (urines, serum, plasma) are selected, along with the required aliquot number. These are editable to the specific situation in case this is required. The freezer position is also opened to allow the specification of where in the local freezer this sample occurs.

4) Manifest sample tracking

Currently only in the PMT application, but with a view to implementation also in the ENS@T-CANCER VRE, is the use of supporting libraries (Apache POI [12]) to upload Excel spreadsheets automatically to the web application. For the PMT study this is important in the uploading of biochemical and sample information that is sent to the central co-ordination center in Dresden. The corresponding manifest spreadsheet file is printed and sent in the shipment of bio-samples, whilst the electronic file is uploaded and input directly into the application.

D. Statistical analysis

An additional requirement to support a distributed biobank is the ability to use the clinical information for statistical output. In this regard, the requirement to translate a relational database structure into a two-dimensional spreadsheet becomes the main challenge. A critical point is to have all the information for a single patient on one line for each. So information captured that exists in a one-to-many relationship – for instance, the biomaterial forms – must be rendered transversely onto one line, with limits of viewing depending on the format used (.csv, .xls, .xlsx).

The use of this formatting style is particularly important when tracking longitudinal information about treatment and follow-up summaries. For the ACC section of the registry, instances of recurrence, surgery resection status and treatments received, can be summarised in a single page.

All of these features are designed with re-usability in mind. The features support requirements common to clinical studies and trials in general so this goes some way to providing a general application template for study platforms.

VI. LESSONS BEING LEARNT IN SECURITY AND ETHICS

Security for any clinical endeavour is a critical consideration, but especially so when the transfer of human biomaterial is involved. The combination of clinical information can enhance the power of possessing knowledge or potentially the material itself if any security breaches occur in either the online or physical world.

To meet the legislative and ethical requirements of an endeavour like this, the ENS@T-CANCER project complies with the directive 02/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications). Additionally, a major achievement in this international project has also been the (necessary) approval by the individual ethics boards of the participant countries, with advice sought from numerous places including the Comité National Informatique et Libertés (France), Garante per la Protezione dei Dati Personali (Italy), der Bundesbeauftragte für den Datenschutz und die Informationsfreiheit (Germany), and in the UK with the Information Commissioner's Office.

Whilst data collected by the participating clinical centres must be shared under the above legal frameworks to select cases and case materials for research by ENS@T-CANCER participants, any research that is carried out on these data or resources is subject to the Declaration of Helsinki, 1964 and the Convention for the Protection of Human Rights and Dignity of the Human Being with regard to the Application of Biology and Medicine: Convention on Human Rights and Biomedicine Oviedo, 4.4.1997 and Strasbourg 25.1.2005 and as such need to be approved by the ethics committees of the clinical centre that contributes the case and the centre that is performing the research study. As access to the ENS@T-CANCER systems often depends on the nature of the study being performed by the investigator, the latter will submit a case or requirements including ethics approval status to the steering committee.

In the technology used, there are several layers of security to be considered in such a complex environment. In the first instance, the security of the applications and their hosted environments needs to be addressed. Adminstrative security is implemented on the virtual machines that host the application. SSH logins are locked down, access by privileged users is restricted with public/private key-pairs, and firewalls are specifically configured in a "default-deny" manner. The application itself is protection and authentication enforced using traditional web application security methods: the application is "gated" using session based tracking (making use of the JSP notion of session rather than interacting with the client using the web security context); the header of each page has a timeout countdown, set to 15 minutes, which logs the user out if the page has been inactive for this long; restrictions are placed on the application input – gathered at each page – which places limitations on parameter inputs (e.g. are numeric parameters input indeed numeric?) and dropping information that is considered dangerous (SQL keywords and potential byte-code characters that are unexpected in context), as well as parameterising the SQL and all inputs. In each of these situations the system is designed to fail in a manner that is as secure as possible ("default-deny").

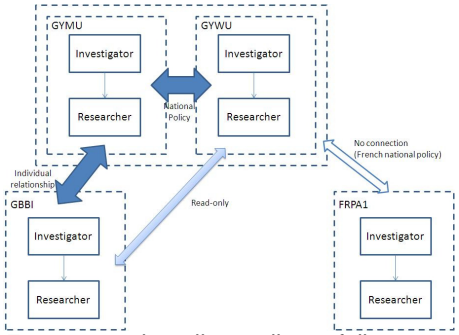

Figure 7: an example policy outline – full access is granted between German centers due to national policy. An individual relationship exists between GYMU and GBBI, otherwise read-only access only exists between GYWU and GBBI. FRPA1 has no connection with the other centers due to the more restrictive French national policy.

Under the directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, users of the ENS@T-CANCER registry have specific privileges depending on their role within the project. Therefore, in addition to the initial technological lines of defense described above it is important that layered inputs of authorization are applied to control inter-center access (it is assumed that whilst the level of trust between consortium members is higher than normal in such an environment, there is still the issue of ownership and protection of the center's data and level of consent), an example shown in figure 7.

The implementation of this type of policy plugs into the business logic of the security module of the ENS@T-CANCER VRE system, the ultimate decision resulting in one of three outcomes: no access, read-only access, write access. This policy and the granularity of access level has been developed allow different views of data and systems within the ENS@T-CANCER VRE. This point was essential since many users only wish to use the ENS@T-CANCER VRE for access within their own centre; or only in their own country; only with the original ENS@T-CANCER project partners, or with any international collaborator involved in ENS@T-CANCER. These policies are similar in many respects to the public management infrastructure developed by the other role-based access control systems [13], where XML-based <role, target, action> models are used to express policies. This information is used to delineate privileges assigned to certain individuals and how these interact (e.g. specific form sections such as "Steroid Metabolomics" are only available to users that possess a "EURINE-ACT participant" role and have ownership of the patient involved). Apache Shiro [14] is currently also being investigated for implementation in this context.

For the purposes of in-depth auditing log4j has been used to track every action of a user: all page accesses and all method calls. The significance of being able to track data provenance is especially important when conducting retrospective monitoring of trials. The next step in this process is to introduce the textual logs into a database so that logs can be organised more logically and searched using more advanced features than string expression matching.

Related to security is the continuity management of the system data. This is achieved by running *mysqldump* through a nightly cron job. The output file is copied to a separate machine as part of that cron job, and every month a copy encrypted using TrueCrypt is created and sent to an offsite location. Scripts to manage backups and file exports are also run to maintain adequate volume space on limited resources. Being able to automate the TrueCrypt process is also desirable (balancing security against the convenience of remembering to perform the procedure). A consideration has also been to use the Amazon Glacier service [15], which provides a cost-effective way of storing large amounts of backup files.

VII. Conclusions

The ENS@T-CANCER VRE is a powerful environment that has galvanized the adrenal tumor research field. The PMT and EURINE-ACT studies demonstrate highly-developed usage of the biobank support features available in the ENS@T registry. The connections and distributed nature between these studies and the central registry illustrate the utility of the environment that grows with each added study, all addressing the four different sections of the ENS@T-CANCER VRE. As the linkage and harmonization of methods grows, this environment will support ever-broader features that can significantly help in advancing the study of these important medical conditions through both technological and professional network development.

A further data challenge not described here is the challenge of "big (biological) data" associated with next generation sequencing technologies, e.g. exome and whole genome sequencing approaches. Hosting and linkage of phenotypic and genotypic data is the focus of the endoVL project, where the specific mutations associated with adrenal tumours are being examined and visualized at the genomic level. Furthermore the consistency of analysis and its translation into a clinical context is also essential, since at present the accuracy of bioinformatics interpretation of physical samples depends greatly on many factors including the applications used for analysis and the preparation of samples. Ensuring the quality and accuracy of bioinformatics analysis in a clinical setting is "obviously" essential.

VIII. Acknowledgements

The work leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 259735 and from the Australian NeCTAR endocrine genomics virtual laboratory.

IX. References

[1] Personalised medicine for cancer: from drug development into clinical practice – Jain, InformaHealthcare, 2005, vol 6, p1463-1476
[2] ENSAT projects – www.ensat.org
[3] The Cancer Biomedical Informatics Grid (caBIG): pioneering an expansive network of information and tools for collaborative cancer research – Kakazu et al, Hawaii Medical Journal, 2004 Sep;63(9):273-5
[4] BioGrid Australia facilitates collaborative medical and bioinformatics research across hospitals and medical research institutes by linking data from diverse disease and data types – Merriel et al, Human Mutation, 2011, vol 32, issue 5
[5] The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine – Elliott et al, Int journal of epidemiology, vol 37, issue 2
[6] Development of an Endocrine Genomics Virtual Research Environment for Australia: Building on Success – Sinnott et al, intl conference on computational science and its application (ICCSA 2013)
[7] ISO 27001 - http://en.wikipedia.org/wiki/ISO_27001
[8] PMT – https://pmt-study.pressor.org
[9] Development of Grid Frameworks for Clinical Trials and Epidemiological Studies – Stell et al, Studies in health technology and informatics, vol 120, 2006, p117-130
[10] Google ZXing project - http://code.google.com/p/zxing/
[11] Interleaved 2 of 5 - http://en.wikipedia.org/wiki/Interleaved_2_of_5
[12] Apache POI – http://poi.apache.org
[13] The PERMIS X.509 role based privilege management infrastructure – Chadwick et al, Proceedings of 7th SACMAT conference, p135-140
[14] Apache Shiro – http://shiro.apache.org
[15] Amazon Glacier – http://aws.amazon.com/glacier

University Library

A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
STELL, ANTHONY; SINNOTT, RICHARD

Title:
e-Enabling international cancer research: lessons being learnt in the ENS@T-CANCER Project

Date:
2013

Citation:
Stell, A., & Sinnott, R. (2013). e-Enabling international cancer research: lessons being learnt in the ENS@T-CANCER Project. In e-Science 2013, Beijing, China.

Publication Status:
In Press

Persistent Link:
http://hdl.handle.net/11343/33370

File Description:
e-Enabling international cancer research: lessons being learnt in the ENS@T-CANCER Project