**Detailed review of Natale et al. 2019 submission to F1000 Research.**
T.W. Clark, Ph.D. 12 November 2019

**General Remarks:**
This article describes the implementation of a major system for capturing and sharing research data at NIH. It is unfortunately not well enough organized for good reader comprehension, and the material presented is not sufficiently well-motivated to understand the system use cases, feature selection, design rationale, and outcomes, which are essential in a publication like this.

The implementation technology used is well-understood, and not the main point of reader interest. The potential value of a report like this is rather to inform us of why the system was built, what gap(s) in needs of researchers it fills, design rationale, implementation experience, who should use it and why, why not use other tools instead, and what the experience in building and using it to date has been - what has been learned, including up-front rationale for selecting the implementation technology and whether this turned out to be a good choice.

As a write-up of a major system in use at NIH, impacting researchers, and asserted to be essential for advancing translational medicine, I would encourage the authors to do a major rewrite of this material, with much greater attention to organization, motivation, and the detailed points outlined below. I'd also propose that success metrics for the system, if any, be discussed, along with any effort to evaluate them.

**Abstract**
"A biomedical research system was developed by combining common data element methodology ..."
- What is "common data element methodology"? Please provide a definition and a reference.

**Introduction**
"Translational Medicine (TM) seeks to develop new treatments for diseases with insights towards the improvement of global health."
- Although literally true, this is not a good enough definition of TM. Britannica says about TM that it "aims to improve human health and longevity by determining the relevance to human disease of novel discoveries in the biological sciences." That is what makes it "translational" – "translating" bench research to patient impact, through pre-clinical and clinical drug development.

"The process of TM is time consuming with translational barriers, from basic research to clinical application, bedside to community use, and from community to policy-making decisions. In overcoming the translational barriers, that biomedical informatics platforms reduce the time it takes for basic research to result in clinical applications. "
- Please define "translational barriers". What does this mean? Provide a reference.
- Statement that informatics platforms reduce latency needs some support - the biggest barriers are attrition in the pharma pipeline (T1) and failures in clinical trials (T2). How do biomedical informatics platforms reduce them?

- Not a rigorous argument. Very general and vague.

**Introduction**, final paragraph
"In this paper we demonstrate the application of CDE concept for developing a Biomedical Research Informatics Computing System (BRICS), providing functionalities that facilitate electronic submission of research data, validation, curation, and archival storage within program specific data repositories. Use of CDEs enhances data quality and consistency within the repositories that are important for advancing clinical and translational research."

- This paragraph should be, or should be a major part of, the article's abstract, with some modification to specify the actual program for which this system and set of CDEs is "program specific". Introduction should then provide the context in terms of supported programs, not the entire general spectrum of Translational Medicine.

**Method**
"The Presentation Layer serves as the secure entry point to the BRICS portal. Various open source technologies and libraries, including Java Server Pages (JSP), jQuery, JavaScript libraries (e.g. such as Backbone.js, Asynchronous JavaScript), and XML are used to make web-pages interactive."

- "Secure entry point". How is security provided?

"This layer also includes Web Start applications"

- "March 2018, Oracle announced it will not include Java Web Start in Java SE 11". - https://en.wikipedia.org/wiki/Java_Web_Start
- Is there a reason to retain WebStart now that Oracle has deprecated it, and how is that to be accomplished?

"To communicate and exchange information between the modules, representational state transfer (RESTful) Web services are used."

- Please reference the 2000 and/or 2002 Fielding REST papers appropriately.

"The Data Layer consists of open source databases such as PostgreSQL, Virtuoso databases, file servers, and data persistence frameworks."

- Virtuoso is not an open source product.
- "Virtuoso databases" - are there several?
- Syntax of this sentence needs attention.

"The Virtuoso database is used to store the data accessed by the Query Tool and to store CDE metadata in the Data Dictionary data."

- Why use an RDF graph database here? Important to know. Not a small point.

"The Repository module uses the PostgreSQL database to store and retrieve data."

- What data goes into PostgreSQL and what into Virtuoso - and why?

"Also utilized are open-source libraries such as Hibernate and Apache Jena for storing and retrieve data from databases."

- The many "such as" statements throughout this article are quite imprecise. Hibernate - what is it and why do you use it? What db is it used with? Likewise Jena, same questions.

"The data layer is supported by the physical infrastructure located within the National Institutes of Health, and is certified at the Federal Information Security Modernization Act (FISMA) Moderate level, conforming to additional USA federal information standards".

- What additional standards?

"To de-identify data, researcher's use the GUID tool"

- No apostrophe in "researchers"
- Why is this done? What is the purpose of the GUID assignment process? Is a mapping of GUID hashes to IDs maintained by the host? Why?

**Information package preparation**

"Researchers are responsible for most of the data submission activities, which includes study FS approval, eForms review, curation, mapping of data elements, and providing associated study documentation."

- Why should researchers submit data here? What is the benefit or incentive? Are they required to? What kinds of data may be submitted? What is the domain coverage of the CDE data dictionary? What is "study FS approval"? "eForms review"? what study documentation is required? Why?

"Two routes of data submission are available for researchers to make data findable... ProFORMS...REDCap".

- Why two? What are the use cases for using one over the other? In particular, why isn't REDCap alone sufficient?
- What is the data submission format? Is it .csv files? XML? If XML, where is the XML schema documented?

"Both routes of data submission validate the submitted data using specific range and values from the data dictionaries for a BRICS instance"

- What is the structure of BRICS data dictionaries? Are they ontologies? Where are they too be found - are the FAIR? How were they constructed?

"Data is stored in its raw form within the repository module in a database that can be accessed by the Query Tool"

- What does "raw form" mean? How is the data stored? are there size or format limitations?

"Access is controlled by a Data Access Committee (DAC) that reviews studies for relevance to a specified BRICS instance (defined by the biomedical program)."

- What studies might or might not be relevant? What biomedical programs are covered?

"By using the repository user interface, researchers can generate digital object identifiers (DOIs) for a study, which can be referenced in research articles."

- Are these Datacite DOIs? What is the interface to Datacite? How is the Datacite or other membership subscription held - does NIH hold the Datacite (or other Registration Authority) membership, or must the user's institution join?

"The repository module provides download statistics for spe- cific studies, enabling the investigator to obtain information on their respective data that has been downloaded for other research activities, and overall increase data sharing and collaboration for additional research goals. Depending on the research studies, BRICS based repositories can host high throughput gene expression, RNA-Seq, SNPs, and sequence variation data sets (Figure 2, stage 3)."

- Who are the current users? How many? what is the study type breakdown? Are these all TBI related studies as implied earlier?

"For the data to be available via the Query Tool module, the raw data is processed through the 'NextGen Connect' tool (integrated interface engine) and Resource Description Framework (RDF) data interchange tool (Figure 2, stage 4)."

- Just dropping in RDF here is very confusing. No discussion of the RDFS or OWL vocabulary. Motivation is required here. Why RDF? Where is the ontology defined? Or is one used? What is the need for inherently graph-structure data, and why an RDF graph?

"The Meta Study module is used for meta-analysis of the data as well as a collaboration tool between scientific groups. "

- How does this module support such activities? Why is it necessary? What does it add to the current state of the art? What is the current state of the art relevant to this module?

"The statistical 'R-box' tool, integrated with BRICS, has been incorporated in the QT, to support analysis without having to download data."

- How? References needed here.

"By default, the user is presented with all studies in the data repository that have data submitted against them. "

- Now looking at Fig 3a the reader can see several PD studies from the Scherzer lab. This is the first time the authors present a domain of discourse for their tool. I would otherwise never have know this product holds PD study data. Much too far into the article to discover this important fact.
- At the same time, this is an article about BRICS. But the heading of the screenshot in Figure 3a says "Parkinson's Disease Biomarkers Program Data Management Resources". There is no mention of BRICS.

"Various datasets (e.g. clinical, cognitive, demographic) are available within the repositories that are integrated to the BRICS instances. "

- Another interesting concept without any motivation, mentioned only in passing. Various repositories are integrated into BRICS instances. Where is this explained?
- The reader must guess that BRICS is intended to support multiple data management repository types, over time, and PD is just the first one.

"Use of CDEs results in making data FAIR by harmonization of clinical, imaging, genomics, laboratory, and biospecimen data."

- Data element harmonization (using CDEs) does not in itself make data FAIR.

"The CDEs are easily accessible from multiple open resources - the PDBP data dictionary, the NINDS CDE project, and the NIH CDE repository. "

- This information does not belong down here. It belongs way up front in the overall description of BRICS.

**Biomedical Program Application**

"The initial deployment of BRICS was to support the U.S. Department of Defense's (DoD) and the National Institute of Neurological Disorders and Strokes (NINDS), FITBIR project. The core functionalities for FITBIR were reusable for developing PDBP, as well other biomedical programs."

"A few highlights of the data repositories resulting from the implementation of BRICS instance for the biomedical programs are provided below."

- This information belongs up front in the intro, or at any rate much earlier in the article.

"The informatics system utilizes the Open Archival Information System (OAIS) model for preserving information for a designated community (group of potential consumers and multiple stakeholders). "

- Reference required here to ISO 14721:2003 or similar. Don't make the reader look for this.

"The implementation of the model highlights the importance of developing Submission, Archival, and Dissemination Information Packages"

- Please explain.

"The informatics software hosted for NTR is within a secure Amazon Web Services cloud platform"

- What cloud services are used and how do they map to the previous Figures? Why is cloud deployment used here?

"Deploying in the cloud environment enhances data access, sharing, and reuse of biomedical research data at larger scale".

- There is no special superiority of cloud over non-cloud deployments for reuse, sharing, or access.
- Deploying in the cloud enhances dynamic scalability. In the environment described, the main utility would be to enable rapid scale-out of new repositories.

- The authors support their statement about advantages of cloud environments by reference 43, an article by the lead author Natale, which actually contradicts their claim on its first page. "Cloud computing services offer secure on-demand storage and analysis and are differentiated from traditional high-performance computing by their rapid availability and scalability of services" - from reference 43 (Navale & Bourne 2018).

"The availability of an automated validation tool with the informatics system makes CDE findable in the Data Dictionary and ensures for data quality and consistency."

- It is not clear why this would be true - please explain.

"Usability of data is enhanced by the adoption of standard imaging formats "

- Is that the only usability-enhancing element in this system?

"The informatics system also supports data discoverability across multiple repositories through the application of the biomedical and healthCAre Data Discovery Index Ecosystem (bioCADDIE)"

- The fact that biocaddie can index data from this system is not a feature of the system, unless special provision is made for it to be indexed. The authors do not describe anything they have done to enable biocaddie to index their data.
- Also: the biocaddie.org website is currently unreachable and does not appear to be maintained.
- The DataMed website for biocaddie's data repository index is reachable, but none of the repositories referenced as BRICS instances can be found there.

**Conclusion**

"Data confidentiality, integrity and accessibility are essential elements of responsible biomedical research data management."

- Nothing has been said in this article about the user authentication and authorization model, which would seem to be involved in confidentiality.