HHS Public Access

Author manuscript

Procedia Comput Sci. Author manuscript; available in PMC 2017 May 03.

Published in final edited form as:

Procedia Comput Sci. 2012 ; 9: 1595–1603. doi:10.1016/j.procs.2012.04.175.

Prototype of Kepler Processing Workflows For Microscopy And Neuroinformatics

V. Astakhov[a], **A. Bandrowski**[b], **A. Gupta**[b], **A.W. Kulungowski**, **J.S. Grethe**[b], **J. Bouwer**[a], **T. Molina**, **V. Rowley**[a], **S. Penticoff**[a], **M. Terada**, **W. Wong**, **H. Hakozaki**[a], **O. Kwon**, **M.E. Martone**[b], and **M. Ellisman**[a]

[a]National Center for Microscopy Imaging Research, Basic Science Building 1000 University of California, San Diego 9500 Gilman Drive La Jolla, CA 92093-0608, USA

[b]Neuroscience Information Framework, Calit2 University of California, San Diego, 9500 Gilman Drive La Jolla, CA 92093-0436, USA

Abstract

We report on progress of employing the Kepler workflow engine to prototype “end-to-end” application integration workflows that concern data coming from microscopes deployed at the National Center for Microscopy Imaging Research (NCMIR). This system is built upon the mature code base of the Cell Centered Database (CCDB) and integrated rule-oriented data system (IRODS) for distributed storage. It provides integration with external projects such as the Whole Brain Catalog (WBC) and Neuroscience Information Framework (NIF), which benefit from NCMIR data. We also report on specific workflows which spawn from main workflows and perform data fusion and orchestration of Web services specific for the NIF project. This “Brain data flow” presents a user with categorized information about sources that have information on various brain regions.

Keywords

Workflow; Kepler; Microscopy; Web Services; Data fusion; Neuroinformatics; Ontology; NCMIR; NIF

1. Introduction

The advent of new large area detectors and high-speed data acquisition technologies in light and electron microscopy are now producing data sets on the order of multiple terabytes [1–3]. With instrument automation, it is now possible to collect million by million pixel images with a data size of just over 2 terabytes. Additionally, with automated montaged tomograms up to 24k × 24k pixels per image, 120 tilts per rotation axis, and up to six rotation axis, the National Center for Microscopy Imaging Research (NCMIR) can produce just under 1 terabyte of raw data. The final computed volumes are approaching 10 terabytes and could

[a]V. Astakhov. Tel.: +1-858-525-5907; astakhov@ncmir.ucsd.edu. [b]A. Bandrowski. Tel.: +1-650-483-0697; abandrowski@ncmir.ucsd.edu.

exceed 100 terabytes for serial section tomography. With data this size, clever workflow, and large memory, multi-node computation clusters must be used for processing, presentation, and visualization of this data. With these large data sets and multiple users on various instruments and processing workstations all trying to access the network disks simultaneously, the technical challenges only multiply.

To accommodate to those challenges we employ Kepler workflow engine [4–5] to prototype NCMIR end-to-end workflows which deals with data coming from NCMIR microscopes. This system is built upon the mature code base of the Cell Centered Database (CCDB) [2], an integrated rule-oriented data system (IRODS) for distributed storage, and also newly developed code to manage the data and metadata from each microscope. It also utilizes service oriented architecture (SOA) to provide integration with external projects such as the Whole Brain Catalog (WBC) and Neuroscience Information Framework (NIF) [6–7], which are benefiting from NCMIR data.

2. NCMIR workflow prototype - main components (CCDB, iRODS, Kepler)

2.1. Cell Centered Database

The Cell Centered Database is used for data flow management. CCDB models the entire process of reconstruction, from specimen preparation to segmentation and analysis. CCDB provides the raw data, reconstruction, and segmented data for download and includes 2D images, animations, and image map visualizations. CCDB was designed around the process of reconstruction from 2D micrographs, capturing key steps in the process from experiment to analysis. CCDB refers to the set of images taken from the microscope as the Microscopy Product. The Microscopy Product refers to a set of related 2D images taken by light (epifluorescence, transmitted light, confocal or multiphoton) or electron microscopy (conventional or high voltage transmission electron microscopy). These image sets may comprise a tilt series, optical section series, through focus series, serial sections, mosaics, time series, or a set of survey sections taken in a single microscopy session that are not related in any systematic way. A given set of data may be more than one product. For example, it is possible for a set of images to be both a mosaic and a tilt series. The Microscopy Product ID serves as the accession number for CCDB. This value was chosen because a given data set may only be taken once, and, therefore, each Microscopy Product represents a unique dataset. All microscopy products must belong to a project and be stored along with key specimen preparation details. Each project receives a unique Project ID that groups together related microscopy products.

2.2. iRODS

i Rule Oriented Data System (iRODS) is a core data management system used by NCMIR to store, maintain, archive, and provide access to the data generated by NCMIR microscopes. iRODS provides NCMIR with the ability to distribute data among available disc space. Such distribution allows parallel access to the data, which is critical in cases of data sets of terra and peta-scales.

2.3. Kepler

The Kepler scientific workflow system is a key component for integration across NCMIR instruments and applications. Kepler was developed at the University of California, San Diego as a scientific workflow engine that combines data and processes into a configurable, structured set of steps that helps implementing semi-automated workflow solutions. Kepler provides a development environment with a graphical user interface for designing workflows composed of a linked set of components called Actors, which can be executed under different Models of Computations. Actors implement specific functions that need to be performed, and communication between actors takes place via tokens that contain both data and messages. A vast array of existing actors greatly speeds up prototyping and development. Models of Computations specify what flows between the actors, how the communication between the actors is achieved, when actors execute, and when the overall workflow can stop execution. From there, the main motivation for using Kepler is the strong need for a standard layer of abstraction that can provide unification among a variety of applications working in real time. The challenge of integration in this computational scientific domain is a highly heterogeneous API for various microscopes. There are no standards among vendors, which makes the task of integration labor intensive. At the same time, there are many different file formats adopted by various vendors in Electron and Optical microscopy. The unique advantage of using Kepler in this domain lies in addressing the key problem of integration by coordination among various applications. The Kepler data-flow model looks applicable to both the NCMIR and NIF projects. Both Centers operate on huge collections of data. NCMIR generates half a terabyte data daily, which needs to be processed on the fly, while NIF serves thousands of requests to neuroscience data collections. Those requests require service orchestration among several data sources. Also, designed workflows can be executed in batch mode from the NCMIR portal. To perform the required levels of coordination, it was found that a PN-process network director was in particular very helpful with the construction of prototypes. Figure 1 illustrates a prototype of “Data flow” that utilizes PN director. Actors presented in Figure 1 encapsulate processes involved in communication with a microscope native API stated in the name of the actor as well as data migration from a microscope to general data storage iRODS. Differences in native APIs are captured by different actors: “4 cam Microscope Workstation,” “1 cam Microscope Workstation,” “Light Microscopy Workstation,” and “SEM,” while data migration components are identical across those actors.

This workflow utilizes iRODS (through “Storage” composite actor) as the data grid middleware system, which provides a uniform interface to heterogeneous data and is available through the new OpenCCDB portal (CCDB portal actor), which provides authentication and authorization. Every piece of data acquired on every microscope is now registered and stored in CCDB. The data acquisition actor was developed to abstract processes involved in data off load from microscope machine to general network storage (iRODS). This actor can be utilized for other applications, such as for microarray coming from different instruments.

The following section provides a list of the current workflow actors.

3. NCMIR workflow prototype - main actors

3.1. “Microscope Workstation” actor

“Microscope Workstation” data acquisition actors encapsulate the behavior of researchers who want to use the microscope. They submit the basic information about the project, experiment, and subject information to CCDB through the instrument user-interface. Then, CCDB generates the unique ID for tracking this data. All of the microscope and detector metadata is extracted using each instrument’s custom APIs. The data and metadata files are automatically moved to temporary network drives in real time during acquisition time using the rsync protocol. Image viewing and annotation services through the Web Image Browser (WIB) are automatically created. Figure 2 illustrates details of implementation for the data acquisition actor, for example, the “4 cam Microscope Workstation” composite actor. Those actors heavily utilize “External Execute” actor as well as “Directory Maker” and “File Writer” actors.

3.2. Storage composite actor

Data migration network performance and network bandwidth are monitored by the management software -“Storage composite actor,” which provides a way to monitor data flow rates from the microscope machines to the network storage.

Those actors also control data validation and archiving. Users log into the portal to find their data easily. From the portal (Fig. 2), users can launch the WIB to visualize their massive data over the Web. Upon approval, the data and metadata are moved to the archival storage in iRODS and a program creates the meaningful folder structure for storing this data. Data is processed by the flow (Fig. 3) through mounting the IRODs managed network drives, and results are synced to the iRODS distributed storage system and registered automatically to CCDB.

3.3. “Cluster” composite actor

The “Cluster” composite actor (Fig. 1) provides an integration of our current workflow with the UCSD Triton Resource, a 256-node, 4 petabyte storage resource, so that portal users can launch their computation-intensive program easily.

3.4. “Integrated Imaging Workflow” actor

The “Integrated Imaging Workflow” actor (Fig. 1) encapsulates various software packages in the processing pipeline. Those packages, such as TxBR, are accessed through a unified protocol for application integration with a Kepler workflow engine. TxBR is an advanced electron tomography code. This code contains alignment, image correction, filtering, and back-projection reconstruction modules.

Special features of TxBR flow include alignment from projected contours of structures in the object, which reduces or eliminates the need for special markers; image and reconstruction dewarping, which increase the accuracy of the reconstruction and allows for high-quality montaging and z-stacking of reconstructions; generalization of the projection model, which permits a wide range of data acquisition protocols; and parallel versions,

which permit the use of the code on parallel machines, computer clusters, and graphics processor boards.

This code is uniquely suited for the processing of wide-field electron microscope images and tilt series, produces graphical diagnostic output at each step, and is almost completely automated beyond the tracking or contour marking steps.

3.5. NIF and WBC composite agent

Finally, we provide external project integration through NIF and WBC agents (Figure 2). Those agents hide integration with projects such as CRBS, WBC, NIF, INCF, CAMERA and can be accessed through REST Web service actors. The usability of some extensions for that end-to-end integration workflow were investigated for the Neuroscience Information Framework (NIF) project. Here we report some progress on that project.

4. NIF Web service orchestration and data fusion by the Kepler workflow engine

One prerequisite of the scientific enterprise consists of searching for effective and useful reagents, neuroanatomy features, identifying genes, transcripts or proteins of interest. This sort of task can be achieved by some sort of text mining in which any scientist should be able to search PubMed or Google for a keyword that recognizes their area of interest and pull back all data that use that particular vocabulary. The Neuroscience Information Framework (NIF) [1–2] performs such a mission to make scientific data and resources discoverable. Another important aspect is data fusion. There are thousands of neuroscience information resources created by a wide range of information providers, including research groups, funding agencies, vendor groups, and public data initiatives that publish information in one form or another. A neuroscience information resource is defined as any electronically accessible site that provides information of interest to neuroscience. It can be a digital library of publications, such as PubMed; it can be a tissue bank; or it can be the Web site of a neuroscience research group that publishes software tools or data sets. Thus, the problem of finding just the right information from one or more sources has not become easier, and it may be very hard for a general neuroscientist to locate a relevant information resource.

The Neuroscience Information Framework (NIF) [1–2] provides a transparent layer to locate and provide access to federated neuroscience information. It hides from the user a number of underlying factors behind the resource finding problem. These factors are not specific to the domain of neuroscience but come into play whenever an information seeker tries to locate resources which have heterogeneous content, provide heterogeneous access mechanisms, and have not been put into a common information framework.

But presenting all information to the user does not seem practical due to the heterogeneity of the content. Some information comes from ontologies and a lot comes from relational (databases) and non-relational sources (texts, Web services). That information needs to be cleaned to remove duplications and discrepancies and then fused and aligned with respect to semantical context. To perform such processing, we are suggesting to use the Kepler

workflow engine [3–4] to help us orchestrate communication among NIF services and also provide a transparent layer for data "fusion" through "Kepler scientific flows."

One of the first workflows designed at NIF to provide an accumulated view on neuroscience data is the "Brain data flow," which presents a user with categorized information about sources which have information on various brain regions. The high level description of the workflow is simple. The user provides a key word which characterizes the area of interest, for example, "brain" or some part of it, such as "cerebellum." The workflow should expand the concept and search for the NIF registered sources that contain anything related to "brain regions." Data are grouped by categories and sources, counts for data records are extracted and a cumulative matrix of categorized data sources with "regions" is returned to the user.

Detailed specification of the workflow requires an orchestration of several Web service calls and some data processing, which are outlined below:

- User provides term "cerebellum" through the NIF user interface, which initiates a call to the workflow server where the workflow gets executed;
- First, the workflow extracts all sources (<federatedResource name="CCDB">) by calling the REST Web service: http://nif-services.neuinfo.org/nif/services/federationSummary
- At the same time, the keyword is treated as an ontological concept and is used to extract the ontological id associated with the concept. That operation gets performed by a Web service call: http://nif-services.neuinfo.org/nif/services/annotate?content=Cerebellum
- Next, a Web service response is parsed to extract values: "Anatomical Structure" and "birnlex_1489" for term "cerebellum" which is translated to the web service call: http://nifservices.neuinfo.org/ontoquest/rel/parts/birnlex_1489?level=3
- A Web service call returns parts of the brain related to initial concept in XML format, which needs to be parsed to extract an "object," such as "Culman" (<object id="1371916-1">Culmen</object>) from XML:

```
<relationship>
<subject id="1371182-1">Cerebellar cortex</subject>
<property id="17089-15">has_proper_part</property>
<object id="1371916-1">Culmen</object>
</relationship>
```

- Extracted values are used to iterate through another Web service call: http://nifservices.neuinfo.org/nif/services/searchSummary?q=Culmen
- Responses are categorized by data type and provide information about data sources (db="SumsDB") and count (<count>43</count>) for data records.
- Finally, information gets aggregated and presented in CVS format for further visualisation by the NIF Web interface:

Cerebellum Culman	Source1 Count
Cerebellum Culman	Source2 Count
.....	
Cerebellum Cerebellar cortex	Source1 Count
Cerebellum Cerebellar cortex	Source2 Count

By using the REST Service actor in combination with some control-flow actors (i.e ArrayToSequence, StringSplitter etc), a functional workflow for accessing and coordinating across various Web services can be built much more quickly then by encoding the workflow by a standard programming language, such as Java or Python.

5. Conclusion

Many of those agents are still under development and undergoing testing, but parts of the system were successfully deployed at the National Center for Microscopy Imaging Research (NCMIR). Those prototypes demonstrate usability of Kepler scientific workflow engine inside the NCMIR and NIF projects for quick prototyping and utilization of system functionality for data management, aggregation, and semantic fusion. Over the process of prototyping end-to-end integration from microscopes to Web environment, we continue to learn more about the workflow engines [8–9]. Also, we have had an opportunity to compare Kepler engine with another one (Mule) which implements Enterprise Service Buss approach. Mule provides a capability to communicate across many protocols but requires Java programming to implement even the simplest logic, while Kepler provides a big library of actors that simplifies the process of development and does not require a specific programming language. At the same time, it was hard to detect when a process finished in Kepler. In our workflows, we often needed to wait until data was written to a file completely before starting the next processing step, and it was difficult to implement such logic in Kepler. To overcome that problem, we wrapped some completed Kepler workflows in a shell script that called Kepler as an external process from a secondary Kepler workflow. It seems that some challenges still need to be addressed in Kepler. It will be helpful to have more actors dealing with various communication protocols. Also, we need to be able to detect whenever a process finishes before triggering another processing step. Thus, the ability to detect the state of an Actor whenever it finishes will benefit many of our workflows.

References

1. Lawrence A, Bouwer JC, Perkins G, Ellisman MH. Transform-based backprojection for volume reconstruction of large format electron microscope tilt series. J Struct Biol. 2006; 154:144–167. [PubMed: 16542854]
2. Martone ME, Tran J, Wong WW, Sargis J, Fong L, Larson S, Lamont SP, Gupta A, Ellisman MH. The Cell Centered Database project: An update on building community resources for managing and sharing 3D imaging data. J Struct Biol. 2008 Mar; 161(3):220–31. Epub 2007 Oct 16. [PubMed: 18054501]
3. Abramson, D., Bethwaite, B., Dinh, M., Enticott, C., Firth, S., Garic, S., Harper, I., Lackmann, M., Nguyen, H., Ramdas, T., Russel, ABM., Schek, S., Vail, M. IEEE e-Science. Oxford, UK: 2009 Dec 9–11th. Virtual Microscopy and Analysis using Scientific Workflows.
4. Altintas, Ilkay, Berkley, Chad, Jaeger, Efrat, Jones, Matthew B., Ludäscher, Bertram, Mock, Steve. Kepler: An Extensible System for Design and Execution of Scientific Workflows. proceedings of

the 16th International Conference on Scientific and Statistical Database Management (SSDBM 2004); p. 423-424.

5. Davidson, Susan B., Boulakia, Sarah Cohen, Eyal, Anat, Ludäscher, Bertram, McPhillips, Timothy M., Bowers, Shawn, Anand, Manish Kumar, Freire, Juliana. Provenance in Scientific Workflow Systems. 2007; 30:44–50. IEEE Data Eng. Bull.
6. Gupta A, Bug W, Marenco L, Qian X, Condit C, Rangarajan A, Müller HM, Miller PL, Sanders B, Grethe JS, Astakhov V, Shepherd G, Sternberg PW, Martone ME. Federated Access to Heterogeneous Information Resources in the Neuroscience Information Framework (NIF). Neuroinformatics. 2008 Sep; 6(3):205–17. Epub 2008 Oct 29. [PubMed: 18958629]
7. Young, L., Vismer, D., McAuliffe, MJ., Tu, SW., Tennakoon, L., Das, AK., Astakhov, V., Gupta, A., Grethe, JS., Martone, ME. Ontology Driven Data Integration for Autism Research. IEEE International Symposium on Computer-Based Medical Systems; 2009;
8. Barseghian, Derik, Altintas, Ilkay, Jones, Matthew B., Crawl, Daniel, Potter, Nathan, Gallagher, James, Cornillon, Peter, Schildhauer, Mark, Borer, Elizabeth T., Seabloom, Eric W., Hosseini, Parviez R. Workflows and extensions to the Kepler scientific workflow system to support environmental sensor data access and analysis. Ecological Informatics. 2010; 5:42–50.
9. Oinn T, Greenwood M, Addis M, Alpdemir N, Ferris J, Glover K, Goble C, Goderis A, Hull D, Marvin D, Li P, Lord P, Pocock M, Senger M, Stevens R, Wipat A, Wroe C. Taverna: lessons in creating a workflow environment for the life sciences. Concurrency and Computation: Practice and Experience. 2006; 18(10):1067–1100.

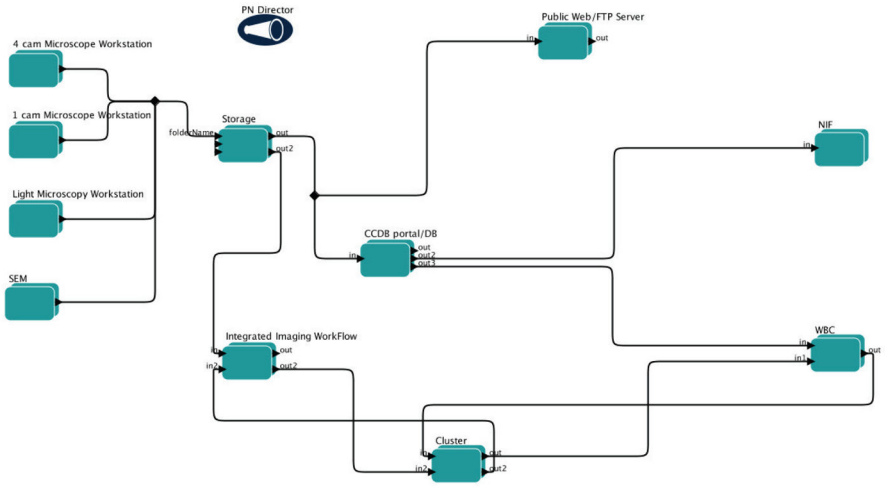

Fig. 1.
NCMIR “end-to-end” application integration flow

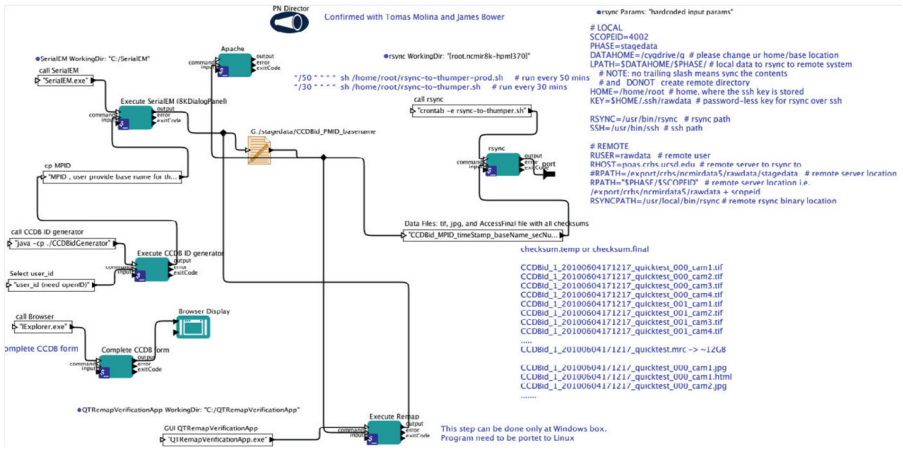

Fig. 2.
Data acquisition actor

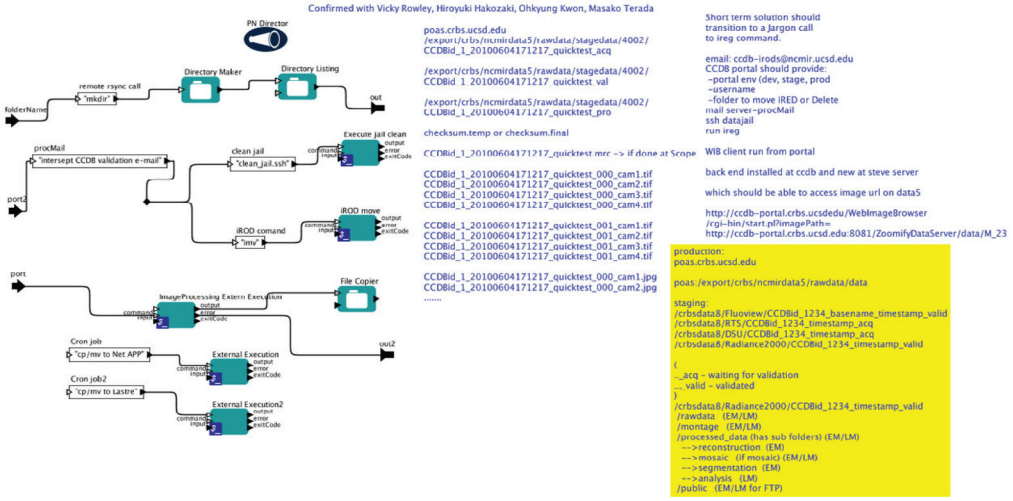

Fig. 3.
"Storage" composite actor

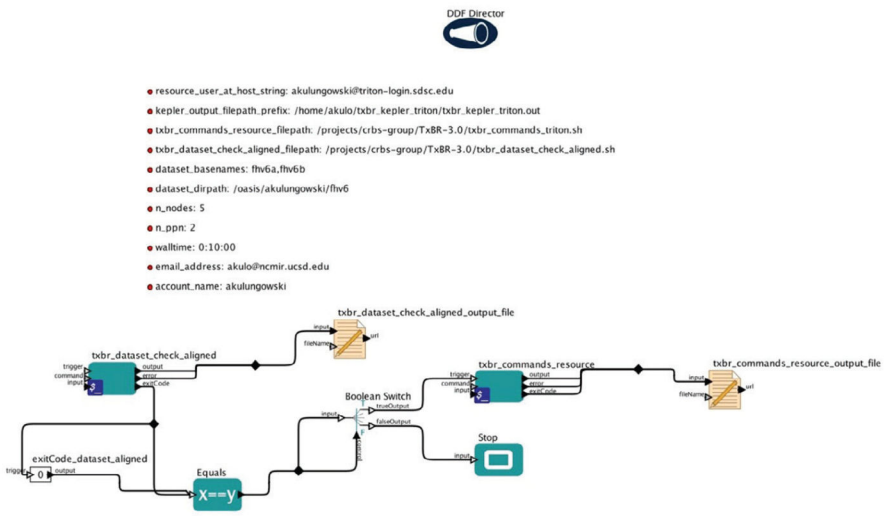

Fig. 4.

TxBR reconstructs 3D volumes by back projecting election trajectories. The TxBR workflow is integrated with the NCMIR system through Kepler.

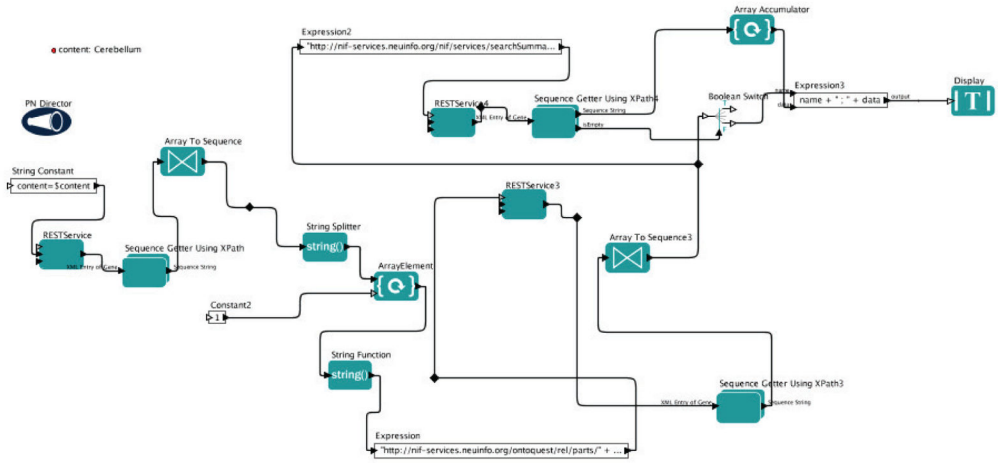

Fig. 5.
NIF “Brain data” workflow with main actors