CHORUS: Collaborative Heterogeneous Observation, Reasoning and Unifying System

1 Vision for the Platform

We propose to coallesce and leverage existing data sets using international standards for data interchange to be shared by several domains of knowledge, including astrophysics, oceanography, remote sensing, fisheries, emergency management and robotics. This platform would allow for the exciting possibility of combining data and algorithms across these different domains. Example use cases include:

1. A forest fire-fighter consulting her smartphone before approaching the fire to see an immersive view of when the fire is at this moment from near real-time remote sensing combined with GIS data of the terrain. Perhaps more importantly, she can also see where the fire will be in five or ten minutes by harnessing simulations of the propagation of the fire using the most recent data on the fire, wind and the forest.

2. The development of a robotic emergency rescue first responder equipment through a simulation of the design in the actual terrain of interest. With parallel computation, thousands of potential designs could be tested and modified using a genetic algorithm in a matter of minutes to find the optimal design.

3. Management of natural resources combined with new techniques such as fracking combined with climate change, potentially taking into account the subtle impact on agriculture at a global scale.

4. A public health researcher correlates medical outcomes in a private database against geographical data to discover new trends underlying disease, without gaining access to any private data.

2 Expected Impacts and Added Value

Many Canadian built research tools either involve domain specific geographical data (e.g. Ocean Networks Canada) or specialized provisioning of distributed computational resources (e.g. CANFAR). Currently, to interact with this data or perform computations, one not only has to be an expert in the particular sub-discipline but also in each interface of the underlying tool or the data representation. We believe these efforts are now mature enough to compose their collective APIs into one standard interface available for users from around the world to collaborate through a web browser.

The key tenet of the proposed platform will be to exploit existing tools, data sets, and standards to provide an interactive interface for industry stakeholders and the public. This will enable multi-disciplinary and multi-sectoral networks of stakeholders to create strategic international partnerships. Not only will exchanges be accelerated, but existing Canadian efforts will be better internationalized. Promotion of standards would dramatically increase access to existing infrastructure and data, foster collaboration among international stakeholders and ultimately yield new understanding of our environment, natural resources, more rapid development of robotic technology and better situational awareness for defense and emergency management.

CHORUS will enable partners to deploy new sophisticated information products take advantage of the new data that will be available in the near future (such as constellations of high resolution SAR and Optical satellites). The enhanced cloud platform will have the capability to perform big data analytics and generate powerful new information products based on the analytics results. Furthermore, we will combine heterogenous remote sensing and in situ data in a uniform interface and catalogue which will for the first time allow global calibration of the remote sensing (i.e. the ground truth), dramatically increasing inherent value of the remote-sensing data and will create an entire new capability of analysing these heterogenously acquired datasets in tandem. We envision the technology to have multiple international market opportunities.

The first is to exploit the powerful new information products that can be generated from big data analytics combined with the huge data archive. These opportunities include creating services around these

information products and/or selling the information products to service companies. Big data analytics is a rapidly growing market segment that is growing at a 26.4% compound annual rate and is projected to be $41.5 B annually by 2018 (around the time this project will be underway and the technology is ready for exploitation). This dwarfs the global earth observation market which is currently around $1.9 B annually and projected to get to $3B annually by 2020. Enhancements made to CHORUS will allow competing directly in the big data analytics market with a powerful cloud based system. Therefore, the market potential for this technology is very high, and it only takes a fraction of a percent of market share to see significant revenues.

Another application of this technology is to take CHORUS and deploy this entire ecosystem on to a private cloud infrastructure that can be sold to customers around the world as a stand alone product. This targets mainly large government organizations and institutions and also large companies (e.g., insurance companies) that are interested in the cloud based computing capability but want to have the system behind their own firewall so they have complete control of the security. The value of this technology is very high to each of these customers (i.e., easily in the 10's of million dollars) which is what it would cost for them to develop this capability plus it would cause them significant delays in acquiring the capability if they developed it themselves. Therefore, a price of several million dollars per instance is readily defendable. Since there are hundreds of organizations around the world that would be highly interested in this type of capability, the revenue potential is greater than $100M.

3 Model for Collaboration

By offering an open and collaborative platform, CHORUS will strengthen, expand and intensify meaningful international collaborations across a variety of knowledge domains, including astrophysics, oceanography, remote sensing, fisheries, natural resources development, energy, agriculture, emergency management, public health, social services and robotics. The need for a common infrastructure, in particular for applications that crosscut both the knowledge domains and datasets involved, will build links and ensure shared benefits between national and international institutions alike. We have successfully assembled leading practioners from both academic and industry stakeholders involved in extracting information from existing Earth Observation platforms as well as databases from observations on the ground. This initial collaboration will identified additional potential stakeholders who will assist the software development tean to address their particular needs.

We will be using several open standards for representation of geographic information, images and distributed computation. These standards are flexible, extensible, well-documented and supported by a large industry performing both defense and industrial simulation. New stakeholders can easily integrate their database in the CHORUS platform by either providing an API that we will adapt for CHORUS or using the open HLA standard for distributed computation. Furthermore, if a new stakeholder wishes to allow model testing on their data without divulging the data itself (perhaps due to privacy concerns), CHORUS would include access to a server that would be restricted to testing models on the private data. By distributing the computation to the data repository, the end user can provide a prospective model to a mutually trusted server that has access to the private database. The server would return a single number or a few numbers to quantify the quality the fit of the model to the data.

The members of the network will meet four times yearly in Vancouver or Victoria in person, but stakeholders will also collaborate informally with the software development team at UVic either through in-person meetings or virtually. The initial group of stakeholders will identify additional partners to understand the detailed needs of the various communities for access and analysis of geographic data (market discovery). The team will develop the specifications required of the platform and in collaboration with the software team create prototypes of the various data models and database access. The collaboration will strive both to identify the particular needs and challenges of developing the CHORUS platform for knowledge translation and sharing (market validation), but also to define these needs and challenges by producing software prototypes to achieve or at least attempt to achieve the goals (productization). For

this to be most productive, the key collaboration will be between stakeholders who outline their needs and the software team that will codify these needs within software prototypes. We believe that SEDRIS standard for GIS data and HLA standard for distributed computation are sufficiently flexible and extensible to meet the needs of various stakeholders but we will also explore additional possible data and computational models in collaboration with our stakeholders at CADC, Urthecast and Magnetar.

4 Strategic Plan

The objective of CHORUS would be to create a collaborative, international development environment for algorithms and tools designed to harvest information from shared earth observation data repositories. Sharing applications within this environment has the following advantages:

1. It allows the system to move the code to the data, which speeds up the processing, and reduces the cost;
2. It will enable model testing on private datasets without divulging private data;
3. It can make the applications available to other clients and users;
4. It encourages collaboration within the CHORUS community, possibly including crowd-sourcing.

Tasks will include:

1. Developing standard workflow practices: for example, users first find data through the use of the catalog API, then assign the selected dataset to a working set, and then launch individual tasks/algorithms or a chained set of tasks/algorithms.
2. Designing automated and efficient mechanisms to deploy algorithms and applications as tasks within the system. These would for example take advantage of Hadoop and Spark for mass batch processing and analysis.
3. Automatically selecting CPU or GPU (CUDA) optimized code for analysis tasks.
4. Exploring strategies such as containerized runtime virtualization (for example using Docker) to run algorithms at scale through the CHORUS environment.
5. Automating how algorithms can be re-factored for speed, for example through feature engineering and with reference to the machine learning literature.
6. Assessing the trade-off space between standardization of workflows and optimization of workflows for data throughput, with reference to various classes of problems.
7. Designing and optimizing the user experience for interacting with and sharing workflow containers within the ecosystem developers and users.
8. Experimenting with “cloud sourcing” of applications. This might for example use a revenue-sharing model similar to Apple’s “App Store”.

The vision for this project is to provide critical new capabilities based on what is now redibly available Earth Observation datasets. The goal is to take CHORUS to a stage where the technology is ready to be commercialized by industrial partners. Modern constellations of commercial surveillance satellites that are now being developed are of great interest to both civilian and defence agencies around the world. Such constellations will consist of many satellites with frequent revisit capability and will generate huge amounts of data. To meet the requirements of many users, those high volumes of data must first be

translated into geometrically and radiometrically corrected imagery, analysed to extract the embedded information, and then delivered around the world in minutes. This capability can only be achieved through Cloud Computing on systems such as CHORUS.

This work builds on existing Geospatial Cloud Platforms and tools that currently enable the ability to generate information products suitable for such tasks as search and rescue and related defense applications that include big data analytics applications. In addition, the IP that is developed will have high commercialization potential and the technology will be taken to a stage that it is ready to be commercialized by industrial partners.

The most critical metric for success will be the engagment factor from the mulitple stakeholders involved. If CHORUS is successful, then it will attract additional datasets, algorithms, and information products from all of the knowledge domains involved. This method of crowd sourcing will also identify internationally-important knowledge gaps and the shared platform will be the foundation upon which we can build new collaborations.

Development of CHORUS will adhere to an Agile Methodology. Agile iterative approach that builds software incrementally from the start of the project, instead of trying to deliver a completed platform near the end of 4 years. In each iteration, stakeholders will be asked to come to a consensus on key elements of functionality to support. This approach will allow the project to most easily adapt to new opportunities and unexpected challenges.

5 Proposed Team

The partnership consists of representatives from UVic, UBC, SFU, Ocean Networks Canada, the Space Telescope Science Institute, International Virtual Observatory Association, and industry partners from IBM, Urthecast and Magnetar Games. The key individuals highlighted in this brief LOI are all Canadians with established extensive international collaborations. Consolidating efforts through CHORUS will intensify these relationships with partners internationally and further expand them to include more knowledge domains.

IBM will provide access to critical cloud-based infrastructure software. UrtheCast will provide multispectral remote sensing data and have strong expertise in GIS. Magnetar games provide expertise on immersive user interfaces and augmented reality. We envision developing web-based prototypes allowing a non-expert citizen scientist to not only access the data but to interact with the data as an expert would, creating new simulations and perhaps more importantly to combine data from different domains to generate new understanding.

Yvonne Coady

Yvonne leads the systems research group at the University of Victoria, exploring cloud-based application infrastructure and scientific visualization. Her group has had impact in optimizations and efficiencies afforded by new hardware and infrastructure in a wide range of scientific applications, including the Thirty Meter Telescope and Near Field Tsunami Detection and Warning Systems. As a co-recipiant of the University of Victoria's Knowledge Mobilization Award, and a co-winner of Johnson & Johnson's Cognition Challenge, she has been both locally and internationally recognized for her participation in Knowledge Translation activites.

Yvonne will manage the network as a whole and as the chief software architect, she manage the software development of CHORUS NCE Knowledge Translation Platform.

Jeremy Heyl

Jeremy is aprofessor at the University of British Columbia and a Canada Research Chair in Neutron Stars and Black Holes. His research has focussed on compact objects, the evolution of stars and galaxies. His discoveries include the first realistic calculations of the mergers of spiral galaxies to form elliptical galaxies, the first theoretical calculation and observational measurement of the evolution of the distribution of the luminosities of galaxies through cosmic time and the first measurement of the diffusion of stars

through a globular cluster. All of these discoveries required the development of new statistical tools and the analysis of large datasets. He is an expert on high-performance computation and the statistical analysis of large datasets.

Jeremy Heyl will engage the core group of stakeholders and grow the network. He will gather the contributions from the individual stakeholders, identifying which databases to include and how to include them. He will also design the initial set of analysis tools available within the platform and will develop the interface in collaboration with partners at STSci and Magnetar Games.

Tania Lado Insua

Tania has investigated environmental impacts of oil-spills on mussel communities using population genetics techniques, and holds a Diploma of Advanced Studies in Marine Biology and Aquaculture from the University of Vigo, in collaboration with the University of Puerto Rico, MayagÃijez. She obtained MS and PhD degrees in Ocean Engineering from the University of Rhode Island with research focusing on applying models of sediment physical properties to past and present climate change. Her most recent research includes international collaborations on diverse research topics such as environmental impact evaluation for renewable energies, paleoceanography, physical properties of the sediment, geohazards monitoring, seafloor observatories and paleoclimate.

Tania will provide expert advice on how include data from in-situ real-time sensors, from Ocean Networks Canada in particular, and from other sources as well. She will also engage other Earth and ocean scientists within Canada and beyond.

Réka Gustafson

Réka is the medical health officer for Vancouver Coastal Health and a clinical assistant professor in the School of Public Health at the University of British Columbia. She works on improving hospital and community physician practice through good science, excellent knowledge translation, unrelenting advocacy, and extraordinary persistence. She has been the voice of reasoned science in CDC projects and programs across the province and continually links the most thoughtful reflection on data with the realities of public health practice and ethics.

Réka will advocate within the CHORUS colloboration for the needs of the public health community and reach out for new stakeholders in this area.

David Schade

David manages the Canadian Astronomy Data Centre (CADC) at the NRC Herzberg Institute of Astrophysics (NRC-HIA) in Victoria. He has worked on the Hubble Space Telescope Medium Deep Survey project and the analysis HST imaging of faint galaxies in the Canada-France Redshift Survey. The CADC is a virtual observatory that archives astrophysical data from a variety of ground-based and space platforms. David is an expert on data archiving from heterogeneous sources and interoperability. He is also a key player in the Canadian Advanced Network for Astronomy Research (CANFAR).

Like geographical simulations, theoretical astrophysics consumes prodigious amounts of HPC resources. And like geographic databases, observational astronomy is data-intensive and requires infrastructure that has not been readily available from organizations like Compute Canada. CANFAR was formed as a partnership of university scientists with the Canadian Astronomy Data Centre which has a 25 year history of providing data management services to the university community. CANFAR is delivering cloud storage and processing services to the data-intensive astronomy community in partnership with Compute Canada and with the support of CANARIE. David will provide key guidance on the design of the CHORUS platform.

6 Management and Governance

Yvonne Coady will be the executive in charge of the entire network: the setting and meeting of short-term goals and the general financial planning. The bulk of the software will be developed at UVic, under her su-

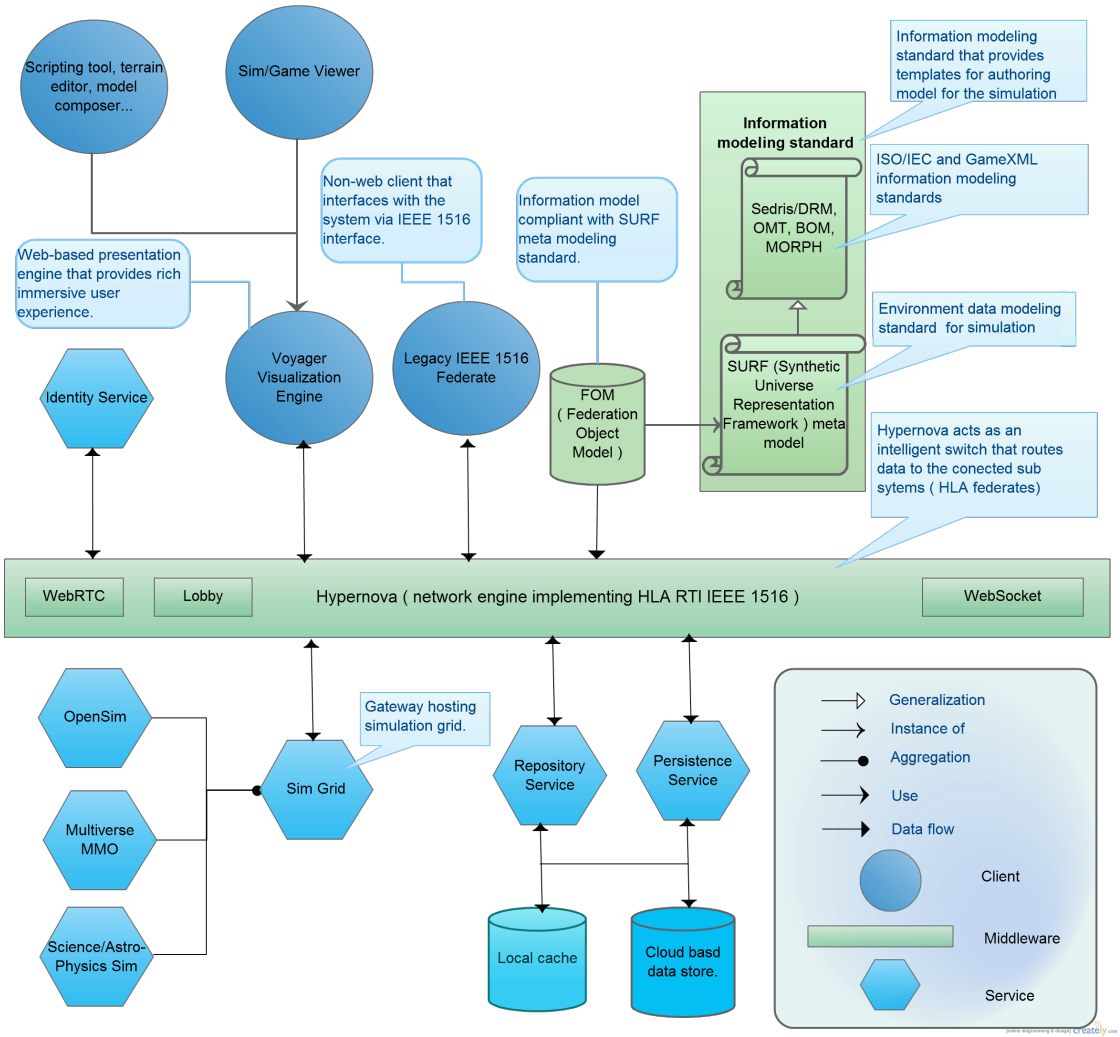
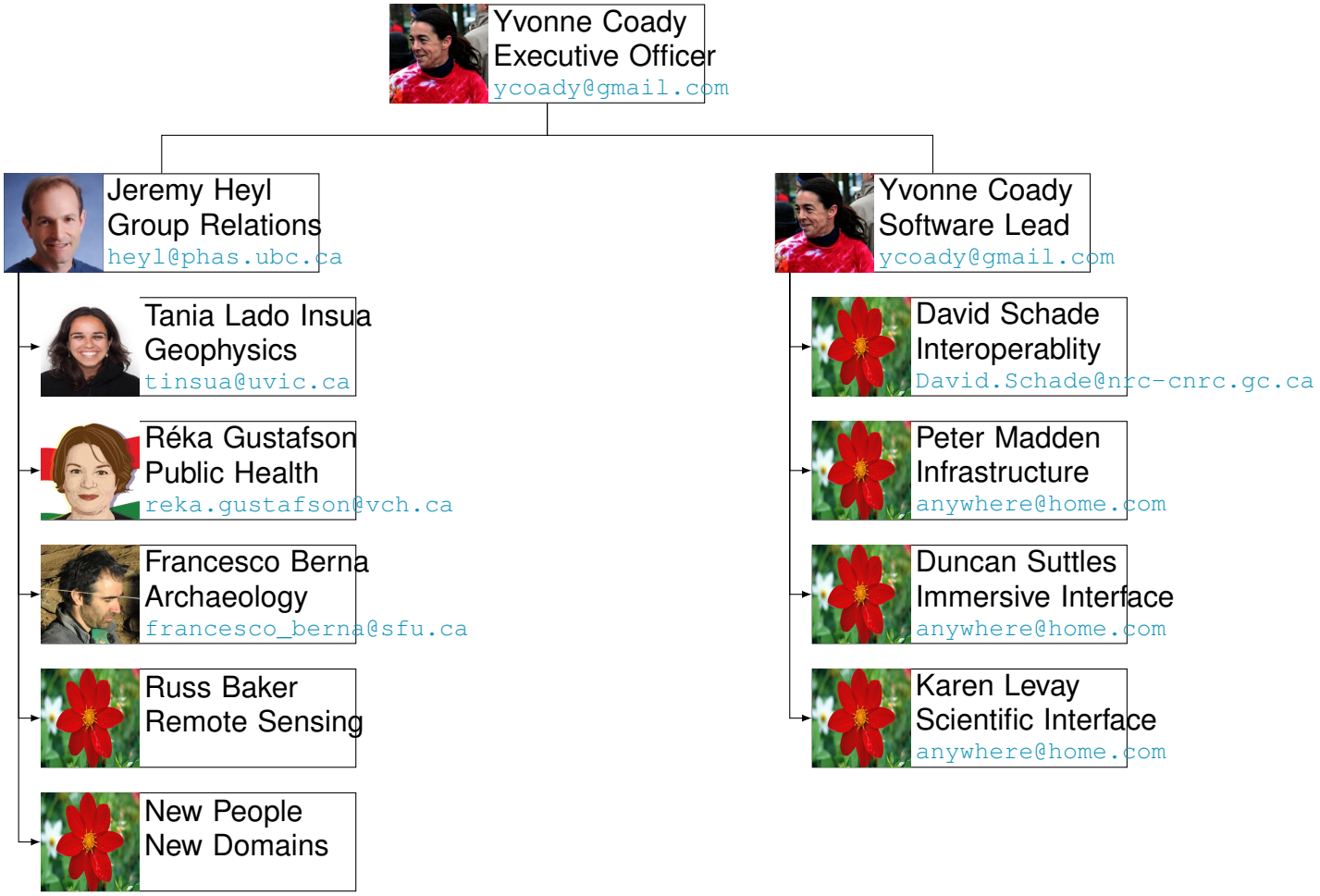


Figure 1: Potential Interoperability Architecture for the CHORUS Software Platform

pervision. She will be in charge of hiring and managing the team to develop the knowledge-transformation platform. She will also design the general architecture of CHORUS.

Jeremy Heyl will engage the core group of stakeholders and grow the network to new domains. The initial domains are remote sensing, geophysics, public health and archaeology. He has already contacted additional potential collaborators in forestry and fisheries, and will develop contacts in other areas of natural resources and social sciences. Jeremy will be in charge of the organization of the formal collaboration meetings as well as the team meetings within the individual domains to discuss their detailed needs and incorporate them into the software development. He will gather the contributions from the individual stakeholders, identifying which databases to include and how to include them. He will also design the initial set of analysis tools available within the platform and will develop the interface in collaboration with partners at STSci and Magnetar Games.

We have identified key stakeholders in several initiual areas. Each of them will identify further stakeholders within their area to discover the market within their area and subsequently they will engage the stakeholders within their area to identify their needs (market validation). They will also interface with the software development team to help create prototypes (productization).


Yvonne Coady
Executive Officer
ycoady@gmail.com
Jeremy Heyl
Group Relations
heyl@phas.ubc.ca
Tania Lado Insua
Geophysics
tinsua@uvic.ca
Réka Gustafson
Public Health
reka.gustafson@vch.ca
Francesco Berna
Archaeology
francesco_berna@sfu.ca
Russ Baker
Remote Sensing
New People
New Domains
Yvonne Coady
Software Lead
ycoady@gmail.com
David Schade
Interoperablity
David.Schade@nrc-cnrc.gc.ca
Peter Madden
Infrastructure
anywhere@home.com
Duncan Suttles
Immersive Interface
anywhere@home.com
Karen Levay
Scientific Interface
anywhere@home.com