PROGRAMMA DI RICERCA STM

Fruitore: ....Lorenzo Bigagli

Istituto di afferenza: .... Istituto sull'Inquinamento Atmosferico (CNR-IIA)

Qualifica……Ricercatore………………………livello……III……………………..

Relazione finale

Nell'ambito del programma di ricerca Short-Term Mobility 2016, ho visitato L'Alfred-Wegener-Institut, Centro Helmholtz per le ricerche marine e polari (AWI), situato a Bremerhaven, nello stato di Brema (Germania), nel periodo 15 agosto-2 settembre 2014, svolgendo attività scientifica sul tema: "Big Open Data in Environmental Research – Open Science for Society".

AWI conduce ricerca in Artide, Antartide e a latitudini temperate. Coordina la ricerca polare in Germania e fornisce logistica e supporto per le spedizioni polari. Ulteriori ambiti di ricerca recente riguardano lo sviluppo di tecnologie e policies per la gestione dati, il cambiamento climatico, l'inquinamento marino, il monitoraggio della biologia marina.

Il referente per AWI in questa attività è stata la Dr.ssa Bernadette Fritzsch, che è stata anche mia co-convener per la sessione "Open Access to Research Data and Public Sector Information towards Open Science" della divisione Earth & Space Informatics (ESSI) della European Geosciences Union (EGU) General Assembly 2016, ossia la principale conferenza europea delle geoscienze. I suoi interessi di ricerca abbracciano le tematiche Open Data e Big Data, nonché la programmazione e ottimizzazione del software scientifico.

Ulteriore contatto è stato il Dr. Hans Pfeiffenberger, che ho conosciuto nell'ambito del progetto FP7 RECODE, sul tema delle policy di Open Access ai dati della ricerca, e ho avuto occasione di visitare nei due anni passati. Il Dr. Pfeiffenberger Dirige il gruppo Infrastruttura IT dell'AWI, è membro dell'Open Access Working Group dell'Helmholtz Association e Chief Editor dell'Earth System Science Data Journal, che pubblica dataset scientifici originali, per promuovere il riuso di dati di qualità nelle Scienze del Sistema Terra.

La visita si è inserita nel mio principale ambito di ricerca attuale, ossia l'investigazione di problematiche e opportunità dell'accesso aperto a grandi masse di dati scientifici, in particolare ambientali (da cui il titolo "Big Open Data in Environmental Research"). Il tema è di grandissima rilevanza per la società e la comunità scientifica europea (cfr. il Pilot on Open Research Data del programma Horizon 2020) e globale (cfr. i GEOSS Data Sharing Principles).

Il mio interesse sul tema si concretizza: nella partecipazione al Gruppo di Lavoro Gestione Dati del Piano Nazionale per la Ricerca nell'Artico/Antartico (PNRA); nei contributi all'implementazione di GEOSS; e, soprattutto, nel coinvolgimento pieno nel progetto FP7 BYTE (Big data roadmap and cross-disciplinarY community for addressing socieTal Externalities). BYTE ha l'obiettivo di identificare linee guida e buone pratiche per minimizzare gli impatti sociali negativi dei big data e massimizzane i positivi, in particolare tramite la costituzione di una comunità sostenibile e multidisciplinare, la BYTE Big Data Community (BBDC), per affrontare le sfide sociali date dalla crescente disponibilità di grandi masse di dati, la cosiddetta “Big Data Revolution”. La BBDC è stata fondata il 1 luglio 2016 e resterà attiva almeno fino al 2020. Io ne coordino l’implementazione per conto di IIA-CNR.

Gli obiettivi della visita scientifica erano: investigare le tematiche Big Data e Open Access in AWI e nei contesti scientifici di interesse comune; coinvolgere AWI nella Big Data Community del progetto FP7 BYTE, di cui coordino lo sviluppo; consolidare ed espandere la collaborazione con AWI, a partire dalle sinergie esistenti, valutando possibili strategie di cooperazione futura.

Ho perseguito questi obiettivi presentando e promuovendo le mie ricerche recenti: esternalità dei Big Data per la società, in particolare nel settore ambientale (FP7 BYTE); Open Access per i dati della ricerca (FP7 RECODE; v. anche il volume da me coedito: “Mobilising Data in a Knowledge Society – Open data movement, ecosystems and data”, la cui pubblicazione è prevista per gennaio 2017); valorizzazione scientifica per lo sviluppo sostenibile tramite l’osservazione della Terra (FP7 EOPOWER).

Anche tramite la somministrazione del questionario riportato in appendice, ho potuto raccogliere informazioni sulle infrastrutture digitali e le politiche di AWI sulle tematiche Big Data e Open Access, relativamente a policy, iniziative, tecnologie, esternalità economiche, legali, sociali, etiche. L’attività mi ha consentito di raccogliere utili feedback sulla visione e sulla roadmap di BYTE e su come costituire, sviluppare, promuovere e sostenere la BBDC (e.g. struttura di governance, criteri di membership, obiettivi, opportunità di finanziamento).

I contatti sviluppati sono stati inclusi nella contact list di BYTE, per coinvolgerli nella BBDC, in particolare sollecitandone la partecipazione al primo Workshop, previsto a Valencia (Spagna) in data 1 dicembre 2016, in concomitanza con il Big Data Value Association Summit. Poiché la BBDC resterà operativa almeno fino al 2020, essa potrà fornire il contesto in cui sviluppare e consolidare gli interscambi tra CNR e AWI, e.g. definendo un accordo-quadro per eventuali future collaborazioni.

Inoltre, come previsto, la visita ci ha permesso anche di collaborare all’organizzazione della prossima edizione della nostra sessione su Open Data/Open Science, nella EGU General Assembly 2017. Con alcune varianti nel titolo, la sessione è attiva nella divisione ESSI dal 2014, con crescente successo (i contributi negli anni sono stati 8, 9 e… 29!), a conferma dell’opinione diffusa che il paradigma Open Research Data stia aprendo una nuova era nella pratica scientifica.

Per il prossimo anno, abbiamo deciso di focalizzarla più sugli aspetti tecnologici e infrastrutturali, strumenti, repositories, DMPs, ecc., mentre, data la trasversalità del tema, abbiamo ideato e proposto un’ulteriore sessione estesa a tutte le divisioni EGU, nel contesto del nuovo raggruppamento inter-divisionale “Interdisciplinary Events”.

Questa nuova sessione afferirà ancora alla divisione ESSI e s’intitolerà “Open Data and Open Science”. Sarà rivolta anche e in particolare ai giovani ricercatori, per la potenzialità di creare sinergie inattese. Sarà focalizzata sulle applicazioni concrete nelle geoscienze degli approcci Open, con success stories, ma anche fallimenti, soluzioni e reti di ricerca. Lo scopo è dimostrare come geoscienziati, cittadini, finanziatori, governi, agenzie e altri stakeholder possano trarre benefici dall’Open Science.

Appendice

Nel seguito, si riportano le risposte consolidate a un questionario somministrato ai ricercatori di AWI, per raccogliere informazioni sulle infrastrutture digitali e le politiche inerenti ai temi Big Data e Open Access. Data la natura prettamente tecnica del contenuto, si preferisce lasciare il testo in lingua inglese.

1 Introduction

This document aims at gathering information on AWI’s e-infrastructure and policies on Big Data and Open Access, particularly in relation to some ongoing initiatives of the geospatial scientific community. It is primarily targeted at AWI researchers involved in infrastructural and technological issues.

In this context, "e-infrastructure" refers to hardware and software assets, including the scientific software and the procedures for management, training and support its continuous operation and evolution.

The following chapters formulate questions to gather general information on AWI activities and perspective on the above topics, as well as more specific questions considering AWI’s role as: producer of research data; data disseminator/curator; and end user of research data.

1.1 General questions

This section addresses AWI’s e-infrastructure and policies in general.

1. What are AWI’s activities on the topics of Open Access and Big Data?
    - Strategic work
        - Work out recommendations, position papers
        - Organize Workshops
        - in Helmholtz
            - Helmholtz Working Group Open Science (Hans Pfeiffenberger – Speaker)
            - Task Group “Access to and Reuse of Scientific Software” constituted by Helmholtz Working Group Open Science (Bernadette Fritzsch)
        - in Alliance of Science Organizations in Germany
            - Working Group Legal Frameworks in the Priority Initiative “Digital Information” (Marcel Brannemann)
            - Ad-hoc Working Group Scientific Software (Bernadette Fritzsch)
    - Adoption of Open Access Policy of the Helmholtz Association at AWI
    - Publication of the ePIC database as an Open Access repository
    - Datenportal Deutsche Meeresforschung/MaNIDA fosters the aspect of workflows and dissemination of Underway data of German Research Vessels in Open Access mode

2. How does AWI contribute to the Ocean Data Interoperability Platform, the Science Europe Research Data working group and the Global Earth Observation System of Systems (GEOSS)?
    - Contributions to ODIP
        - AWI is associated partner in ODIP
        - Connected with work in RDA
        - MANIDA (see https://epic.awi.de/31897/1/ODIP_Macario_final.pdf)

    - Sensor data, cruise summary reports (see http://ioc-unesco.org/index.php?option=com_oe&task=viewDocumentRecord&docID=16162)
- Contributions to GEOSS
    - PANGAEA (OAI-PMH)
    - Friedrich-Hustedt-centre for diatom research (see https://web-apps.awi.de/Hustedt-Diatoms)
    - Taxonomic and biogeographic information on plankton (see http://planktonnet.awi.de/)
    - WRMC: World Radiation Monitoring Centre (see http://www.bsrn.awi.de/)

3. The FP7 project RECODE has identified several barriers that hinder Open Access to research data: cultural barriers, technical barriers, legal barriers, ethical barriers, financial barriers, political barriers. In AWI's experience, how would you rank their importance?
    - Ranking depends on personal point of view
    - Low ranking for ethical barriers
    - High ranking for political, cultural and technical barriers

4. The FP7 project BYTE is addressing the many social externalities of Big Data, particularly in the environmental sector. What is AWI's perception of the implications of collecting, linking and analyzing huge amounts of data, in relation to societal issues (consequences on privacy, economic development, etc.)?
    - New (additional) themes in data collections triggered by public discussions. Examples:
        - collection of data about pollution of oceans by litter and microplastics → LITTERBASE will go online in early 2017
        - Retreat of sea ice → sea ice portal provides daily updated ice charts of the Arctic and Antarctic (see http://www.meereisportal.de/)

5. The FP7 project EOPOWER has worked to create conditions for sustainable economic development through the increased use of Earth observation products and services for environmental applications. What is the perspective of AWI about the role of Science (namely Earth Observation) to make Society better?
    - Advisory function for political decision makers and society
        - participation in IPCC Assessment Reports with several contributions and in leading positions (e.g. IPCC work group II Technical Support Unit stationed at AWI, head Hans-Otto Pörtner as co-chairperson of wg II)
        - administrative office of the German Advisory Council on Global Change (WBGU) is affiliated with the AWI
    - Contributions in the development and operation (training) of the early warning system for Tsunamis in Indonesia (group is part of the computing centre)
    - Transfer of research results into society
        - Climate Office for polar regions and sea level rise
            - Mission: "is the transfer of knowledge into the society and the dialogue with them to improve the social impact of the scientific results."
            - Meereisportal.de/seaiceportal.de as a platform for providing information about sea ice for interested society and data for scientists as well
        - AWI North Sea Office
            - Communicates scientific knowledge about North Sea to policy, conservation agencies and the public
            - Developing strategies for sustainable management
    - Coordination of Helmholtz climate initiative REKLIM at AWI
        - Provide a basis for climate related decision support

- AWI represents Germany in consulting and organizing the process for the Scientific Committee to provide the scientific basis for the evaluation of marine protected areas (MPAs) in the Weddell Sea for CCAMLR, Commission for the Conservation of Antarctic Marine Living Resources.

1.2 Producer viewpoint

This section addresses AWI as a producer of research data, e.g. e.g. conducting campaigns that generate new raw data.

1. What and how many research data are produced by AWI?
   - ~50 TB/year in SAMFS in-house data archive
   - Overall amount of data which is produced, but not archived cannot be estimated
   - Measurements, campaigns, satellite data, model data

2. What percentage of that data is released with an Open Access policy?
   - Estimated 20% (only within special projects with their own Open Access policy)
   - In the future, it will be more, in connection with the adopted Open Access policy of the Helmholtz Association

3. What metadata are in use at AWI to describe data, particularly their quality?
   - PANGAEA metadata model
     - ISO19115/139 compliant
     - components: Project, Campaign, Event, Data
     - Quality flags for Near-real-time Data
   - Quality control during
     - Publication process (Scientists)
     - data ingest process (Data curators)
   - Efforts to establish concept for assuring data quality for Underway and monitoring data in MANIDA (Marine Network for Integrated Data Access)

1.3 Disseminator/Curator viewpoint

This section includes technical questions addressing AWI as a disseminator and curator of research data, in charge of an e-infrastructure for storage and access to research data.

1. Can you describe AWI e-infrastructure and policies for Open Access and Big Data analysis?
   - Data Storage
     - Data Archiving System: Hierarchical Storage Management with Cache (~300 TB) and tapes (capacity ~2,3 PB); filesystems mountable on several compute platforms for analysis
     - Several fileservers in working groups
   - Compute Server
     - CRAY CS400, 11.232 compute cores/308 compute nodes (Xeon Broadwell), Fast parallel filesystem
     - NEC SX-ACE, 32 compute nodes
     - Several compute servers on group level
   - Policy – adopted Open Access policy of the Helmholtz Association, but up to now no controlling instances are in place at AWI (see http://os.helmholtz.de/open-science-in-the-helmholtz-association/)

2. What is AWI data management policy (e.g. on publication, online access, preservation, curation)?
   - 3 categories of data
     - Individual data
     - Group data for collaboration
     - Permanent archive or irrecoverable data (includes PANGAEA)
   - Provides data in several specialized data portals (e.g. PANGAEA, planktonnet)
   - Data Scientist – concept developed during the last two years, implementation phase started → Data Scientists in the research working group as person in charge for data management and archiving, as well as participating regularly at an overarching educational curriculum, to gain and exchange data science and data management skills in accordance to the specific needs of the institute, the science community and the own working group

3. Does AWI support data journal, e.g. ESSD? If so, how?
   - Hans Pfeiffenberger, Hannes Grobe, Gert König-Langlo – Editors
   - Experiences see http://epic.awi.de/34846/1/2014-01-29-Pfeiffenberger-APE2014-print.pdf
   - Scientist are encouraged to publish their data in ESSD →our publication data base shows about 30 publications in ESSD

4. Does AWI support the Linked Open Data approach to associate research data to additional complementary information? E.g., raw data, ancillary data (metadata), scientific publications, supplemental information (news articles, multimedia, etc.)
   - PANGAEA Linked Data – Collaboration with Elsevier, link between articles and data and SCHOLIX, a framework for Scholarly Link eXchange (http://www.scholix.org/)
   - Links between articles and data in our internal publication repository (see http://epic.awi.de)

5. What solutions are in use at AWI for searching and identifying data? Please reference specific standards and technologies for metadata management and cataloguing.
   - Cataloguing based on PANGAEA
   - full text search engine based on Apache Lucene
   - Standards: Metadata ISO 19115 compliant, can be harvested via various protocols (OAI-PMH, Catalog Service for the Web)
   - OGC-Standards: SOS, WFS, WMS, SensorML
   - ODV – Ocean Data View implementations use SeaDataNet/NERC standards of Quality Flags and Vocabularies

6. What solutions are in use at AWI for access control and auditing data usage?
   - Google Analytics
   - Piwik Web Analytics

7. What solutions are in use at AWI to manage and optimize the energy footprint of the e-infrastructure?
   - Besides general efforts to optimize energy consumption of the IT infrastructure, no specific, additional efforts to minimize the energy footprint of the e-infrastructure are in place.

1.4 End user viewpoint

This section includes technical questions addressing AWI as an end user of research data, extracting knowledge to support the industry, governmental agencies, etc.

1. What field of science do AWI researchers work in?
   - Polar and marine research → Climate Science, Oceanography, Meteorology, Glaciology, Biology (Biodiversity, Genomics, etc.)

2. Apart from AWI's data, what other data do they need to effectively carry out their research work?
    - Data from many other research institutions worldwide, e.g.
        - Space and air born radar and image data
        - Simulation data for model intercomparison projects (e.g. via ESGF)

3. Does AWI take user feedback into account for improving the quality of data?
    - Special User feedback category at the website
        - PANGAEA Ticketing system
        - MANIDA Feedback box → data provider (no control or execution mechanism is implemented!)
        - Individual contact with principal investigators and data providers.