Using Metadata, Data/Service Quality and Knowledge to Facilitate Better Data Discovery, Access, and Utilization for Supporting EarthCube

Chaowei Yang (GMU CISC), Myra Bambacus (NASA), Karl Benedict (UNM), Doug Nebert (FGDC), Doug Mochuney and Sue Hazlett (USGS), Paul Houser (GMU CISC), Robert Raskin (JPL), Yan Xu and Daniel Fay (Microsoft), Abdelmounaam Rezgui, Qunying Huang and Chen Xu (GMU CISC)

1. Introduction

We have been collecting large amounts data about our home planet from the upper atmosphere (such as ozone change) to well beneath the surface (such as the ocean floor and deep wells). The data record of the Earth in its spatiotemporal context is of great value in understanding the past of our Earth system and to predict the future by analyzing change (Donner 2009). A variety of services and procedures have been built to disseminate, analyze, and utilize data through online tools. Relevant cyberinfrastructures (CI) are deployed to support access to services by relevant disciplines, agencies, and countries. All the data, service, and CI resources built to share and use the Earth data are critical to solve 21st century challenges of geosciences.

Globalization requires sharing and coordinating physical and social resources across the planet (NRC 2003). The globalization also bring many grand challenges, for example, 1) climate is changing and sea level rise is occurring with more polar/glacier ice melt (NRC 2009b; Yang, Nebert, and Fraser, 2011c); 2) severe weather and disasters, such as hurricanes and tsunami, is more frequent (NRC 2011); 3) global social activities speed up the transmission and outbreak of contagious diseases (NRC, 2009a, such as SARS and H1N1); 4) the desertification of land surface, deforestation, and fresh water shortages are becoming more severe . To better address these global challenges, the data about our planet should be shared across domains and across jurisdiction boundaries to enable global geoscience research and emergency response.

The NSF (2009b) CI vision and NSF (2009a) geoscience vision highlight the need for Earth system monitoring; EarthCube (NSF 2011) is proposed at the right time for integrating the resources required to address the grand challenges. Metadata and semantics for describing the resources are central to this integration process to enable users to discover and evaluate the information resources that exist for Earth science studies and applications. To enable across domain and jurisdiction boundary sharing, the standardization (by FGDC, OGC, ISO, ANSI and others) of metadata is critical for identifying interoperable resources and accessing and using these data on the fly. The GEOSS (Global Earth Observation System of Systems) Clearinghouse provides an exemplar global system to harvest and search the distributed metadata collections using standardized interfaces and formats/contents. Data and service quality also plays a significant role to help end users choose the resources that best fit their research, application, or educational needs. The ontology and semantics linked (directly or indirectly) to the resources can help improve the discovery process with better accuracy (Zhang et al., 2010). To address millions of concurrent users when EarthCube becomes operational, support for a spatiotemporal metadata structure, index, ranking, and searching will be crucial. This white paper synthesizes a number of collaborative research and development projects across a variety of agencies, organizations, and companies to demonstrate the utilization of metadata, semantics, and quality to better discover, access, and utilize geoscience resources.

2. Technical Challenges for Sharing Geospatial Resources

The grand vision of EarthCube to integrate geospatial resources (individual subsystems including biosphere, atmosphere, lithosphere, and social and economic systems) at various temporal and spatial scales to support 21st century Earth science and address the challenges (NSF2009a; Yang et al., 2010) requires several technical problems to be thoroughly addressed.

Geographical Distribution and Heterogeneity: Geospatial resources (OGC 1998; Yang et al., 2011) are: 1) globally distributed where multidimensional and massive observations, model outputs, and other resources are distributed across states, countries, and continents; and 2) heterogeneous in that the geospatial data represent different Earth phenomena using different conceptualizations, representations, models, and formats.

Computing Intensity: The analysis and simulation of Earth science phenomena are extremely computing demanding. For example, Earth science phenomena correlation analysis is computationally expensive (Kumar, 2007). The periodic-like phenomena simulation in the Earth system requires the iteration of the same set of intensive computations over time. HPC is usually adopted to speed up the computing process, and often, the computing demands exceed the computing capacity and become a driver for computing science advancements (NRC 2010; Yang et al., 2011).

Spatiotemporal Intensity: Most geospatial datasets are recorded in space-time dimensions either with static spatial information at a specific time stamp, or with changing time and spatial coverage (Terrenghi et al., 2010). The advancement of sensing technologies increases our capability to measure more accurately and obtain better spatial coverage in a more timely fashion (Goodchild, 2007). The spatiotemporal nature of the datasets and their supported sciences make it critical to understand the resources in the spatiotemporal or dynamic manner (Hornsby and Yuan 2008; Yang et al., 2011).

Concurrent Intensive: The popularization of geospatial systems and EarthCube will attract millions of concurrent accesses for real time or near real time data for applications such as real time traffic routing (Cao 2007) and emergency response. This situation will pose great opportunities and grand challenges to relevant scientific and technological domains, such as broadband and cluster computing, privacy, security, and reliability of the information and systems, as well as many other challenges facing massive user systems (Brooks et al., 2004).

Interdisciplinary Collaboration: To better address 21st century Earth science and application challenges, crosscutting solutions are required to integrate interdisciplinary data, information, and knowledge for science advancements (Horvitz and Mitchell 2011). For example, NASA's Joint Agency Committee on Imagery Evaluation and the Commercial Remote Sensing Policy Working Group collaborated with NIMA, USGS, NOAA, and USDA to verify/validate commercial remote sensing sources/products for Earth science research[1]. Such interdisciplinary collaborations pose governance and management challenges for the EarthCube (NSF 2009a and 2009b; Mitchell 2011; Yang et al., 2011a).

NSF Past GCI Assets and Initiatives: The NSF funded several large geoscience cyberinfrastructure initiatives in the past decade to address the Earth's complex systems and to better understand the planet as a whole. Examples include: GEON[2] which is a collaborative project that develops cyberinfrastructure for integration of 3- and 4-dimensional Earth science data ; Unidata[3] : whose objective is to provide data services, tools, and cyberinfrastructure leadership that advance Earth-system science, enhance educational opportunities, and broaden participation; CUAHSI[4] (Consortium of Universities for the Advancement of Hydrologic Science, Inc.) that enables the university water science community to advance understanding of the central role of water to life, Earth, and society; NCAR[5] (National Center for Atmospheric Research) devoted to service, research and education in the atmospheric and related sciences; NEON[6] (National Ecological Observatory Network) which collects data across the United States on the impacts of climate change, land use change and invasive species on natural resources and biodiversity; DataONE[7] (Data Observation Network for Earth) which focuses on infrastructure development (consisting of expertise, technology, and standards) for environmental data and tools; the OOI[8] (Ocean Observatories Initiative) program which is founded upon the goal of sustained ocean measurements contributing to Earth-ocean-atmosphere dynamics; and LOOKING[9] (Laboratory for the Ocean Observatory Knowledge INtegration Grid) which researches the identification, synthesis, and assemblage of existing and emerging concepts and technologies into a coherent viable cyberinfrastructure design.

3. An Integrated Data Discovery, Access, and Utilization Approach

To address the technical challenges for integrating the geospatial resources to support better Earth science and applications in the 21st century, **we collaboratively** conducted over 30 projects **among** a number of **partners** including NSF, NASA, USGS, EPA, NPS, Microsoft, Intergraph, and International Visualization Systems. The vision of EarthCube drives our efforts as we integrate the technologies developed through the projects to support the discovery, access, and utilization of the Earth science resources.

Framework: Our research and development have focused on metadata, catalogs, clearinghouse, spatial web portal, service quality, semantic, and visualization to utilize distributed computing to enable efficient data

[1] http://www.nasa.gov/pdf/55397main_14%20ESA.pdf
[2] http://www.geongrid.org
[3] http://www.unidata.ucar.edu
[4] http://www.cuahsi.org
[5] http://ncar.ucar.edu
[6] http://www.neoninc.org
[7] https://www.dataone.org/about
[8] http://www.oceanobservatories.org/
[9] http://www.calit2.net/research/areas/environment/project?id=62

discovery, access, and utilization (Fig. 1). Through the portals, users can request the core services that can harness vast geospatial resources, which are registered by contributors or resource owners with descriptive metadata through the portal. User requests trigger relevant core components to best handle the requests. The system takes a service-oriented architecture (SOA) design to hide from the users the underlying complexity, e.g., data accesses, processing, and job scheduling.

Components: The core components of the framework (Fig. 1) are given in Table 1.

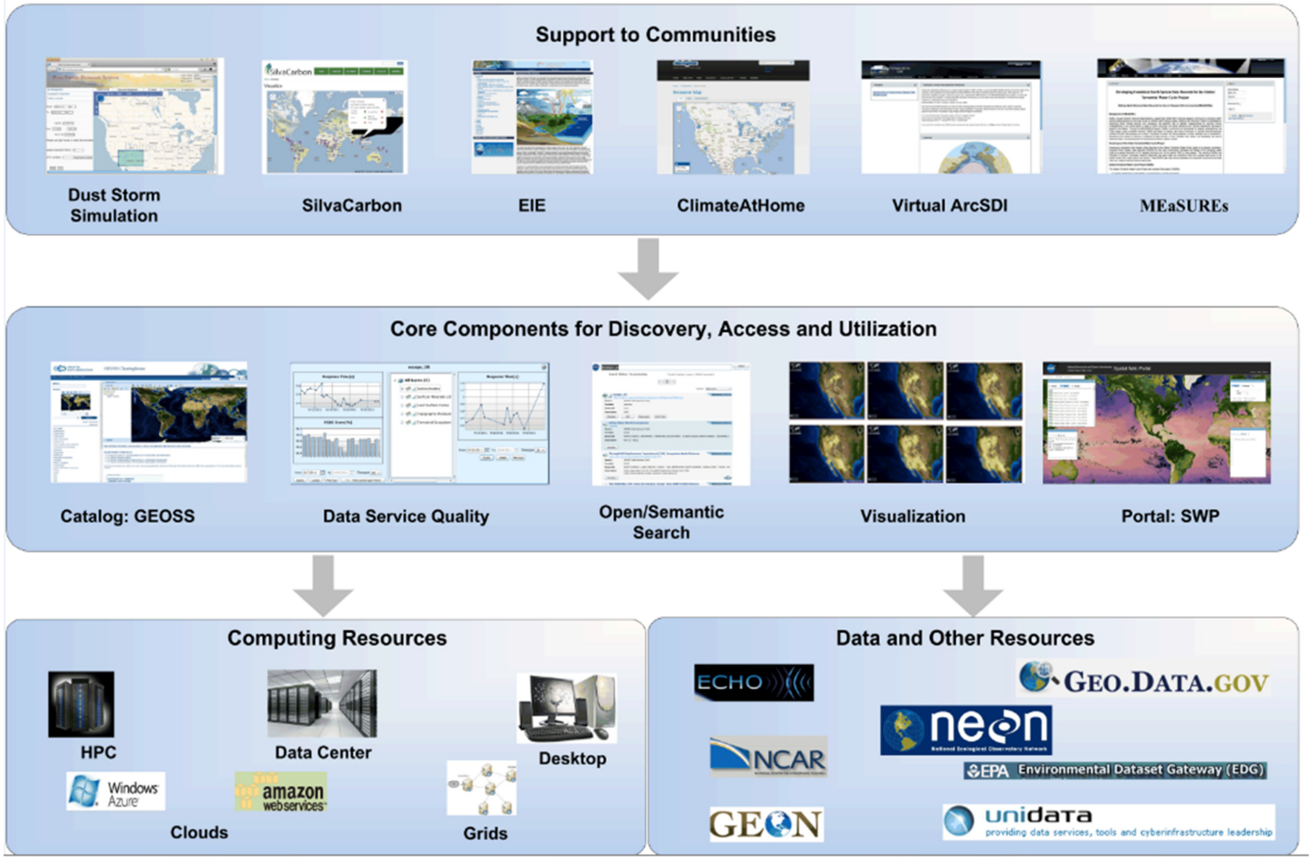

Figure 1. System Architecture

Component	Capabilities
GEOSS Clearinghouse	• Metadata management using ISO-19139 and FGDC CSDGM standards • Catalog harvesting based on OGC-CSW (Nebert and Whiteside 2005), Z39.50, OAI-PMH and WAF protocols. • High performance text and spatial indexing for metadata search • Integrated viewer to visualize Web services content • Scheduled harvesting and synchronization of metadata • Publishing of metadata with local interface and remote search API
Spatial Web Portal	• Search for multiple resources, e.g., services, computing, data, model, etc. • Search resources using multiple search modes, e.g., spatial, temporal, or semantic. • Use of resources based on their status, e.g., service quality, geographic location, etc. • Access spatial service resources in a direct, fast fashion. • Visual delivery through scientific visualization using 2D/3D/4D visualization.
Service Quality	• Assist users to identify best services among many candidates services and also provide a mechanism to help service providers find potential bottlenecks and quality problems. • Provide functions to validate, score and rank geospatial web services including OGC WMS, WFS, WCS, CSW, WPS, etc. Support finer-grained monitoring and evaluation capacity that operates on each web service's operation • Different granularity from service, dataset to data item. • Provide standard web services API to add new monitoring and also retrieve services performance historical information and comprehensive evaluation results.

	• Visualize and analyze functions based on statistical results. • Provide theoretical and information foundation for QoS-aware optimization and planning for web services composition.
Open and Semantic Search	• Distributed search: multiple search sources (GEOSS CLH, GOS, CSR, and GCMD, etc.) with various access protocols or APIs (e.g., CSW, Z39.50, OpenSearch, RSS, ATOM) • Support heterogeneous metadata (e.g., ISO 19115/19139, FGDC CSDGM, DIF, SERF, ebRIM, RDF, SKOS, OWL) • Data crawler functionalities to support Earth science data discovery • Knowledge-based reasoning with Earth science ontology (SWEET, Raskin and Pan 2005) • Ontology-aided search to improve the search prevision and recall, e.g., schema.org • Results ranking based on semantic similarity evaluation
Multidimensional Visualization	• Online visualization tools to facilitate geographic featured information visualization • Scientific 3D/4D/5D visualization to communicate dynamic Earth phenomena • High performance computing enabled rendering capabilities for large scale Earth science data • Interactive visual analytical functionalities to support geospatial analysis • Collaborative visualization of data and resources over the Internet
Computing Resources	• The computing platform to support the processes of data discovery, access and utilization, and model simulation and prediction vary from different types based on the processes requirements. • Most data and search services are hosted on GMU CISC servers. • Model simulation and prediction, such as dust storm simulation, usually adopt traditional cluster HPC resources, or grid computing platform if no or seldom communication is required for the model. • Cloud computing platforms, including Microsoft Azure[10] and Amazon EC2[11], have also been employed for web applications, such as GEOSS Clearinghouse and Spatial Web Portal.
Data Resources	Data from a variety of organizations and systems are harvested or can be searched in a distributed fashion: • NOAA National Climatic Data Center (NCDC)[12] • EOS Clearing House (ECHO)[13] • FGDC Geospatial-One-Stop(GOS) and its successor geo.data.gov[14] • NASA Global Change Master Directory (GCMD)[15] • NASA ESG Open standards and protocols are used to access and integrate data from different organizations, locations, and with different types of data. Also integrated to the system are the following protocols: OpenDAP/THREDDS Catalogue, OAI-PMH, ISO 23950 "SRU", and GeoNetwork "native".

Table 1. Components and Their Capabilities

4 Support to Communities

The core components and services developed support multiple Earth science applications across agencies and organizational communities, such as NASA, USGS, FGDC, NPS, and ESIP. These applications include:

- **Dust Storm Forecasting**[16] predicts dust storm for public health applications in collaboration with the Univ. of Arizona and the Univ. of New Mexico**,** NASA**,** Microsoft and NSF.
- **GEOSS Clearinghouse[17]** provides a search capability against existing catalogues from GEO members and participating organizations.
- **SilvaCarbon Portal**[18] is a USGS led multi-agency project and supports the SilvaCarbon program, part of the GEO Forest Carbon Tracking task, a component of the intergovernmental Global Earth Observation.

[10] http://www.microsoft.com/windowsazure/
[11] http://aws.amazon.com/ec2/
[12] http://www.ncdc.noaa.gov/oa/ncdc.html
[13] http://gcmd.nasa.gov/records/EOSDIS-ECHO.html
[14] http://geo.data.gov/geoportal/catalog/main/home.page
[15] http://www.esipfed.org/content/nasas-global-change-master-directory-gcmd-releases-new-software-version
[16] http://cischpc.gmu.edu/
[17] http://clearinghouse.cisc.gmu.edu/geonetwork
[18] http://swp.gmu.edu/silvacarbon

- **ESIP EIE**[19] is an interoperable and integrative Earth science data management platform proposed by the Earth Science Information Partnership (ESIP) to provide a solution for addressing this challenge.
- **Climate@Home**[20] is a NASA project to develop a virtual supercomputer by leveraging citizens' idle CPU cycles to advance climate studies and other Earth and space sciences.
- **Arctic SDI**[21] is Virtual Arctic Spatial Data Infrastructure, a USGS project to conduct a survey of available Web Map Services about the Arctic region and to prototype a viewer application that would display the multiple WMS from where they are originally hosted.
- **Water Cycle Project**[22] is a NASA project that aims to develop consistent Earth system data records (ESDR) for the global terrestrial water cycle at a spatial resolution of 0.5 degrees (lat-long) and for the period 1950 to near-present.

5. Future Directions

NSF (2011) envisioned EarthCube as the convergence towards an integrated system to access, analyze and share information that is used by the entire geosciences community. The construction of a CI to enable this goal poses grand challenges in many aspects including the integration of currently isolated CIs from multiple domains, the evolution of GCI from a technology-centered to a human-centered paradigm, the advancement of CI to support multiple science domains by simulating complex phenomena in a virtual fashion, and the acceptance of CI by a broad range of stakeholders who use geospatial data. With the advancement of technologies, more study in social sciences is urgently needed to enable efficient governance and collaboration among team members, communities, and domains (Yang et al., 2010; Poore, 2011).

Developing the future CI for the geosciences will require research and development on multiple fronts: 1) work must be done to improve resource discovery (e.g., data, services, etc.) and knowledge sharing, 2) better tools to explore the discovered resources must be developed, including tools for modeling, simulation, analysis, and visualization, 3) improved methods for generating and capturing metadata as a parallel process with data product generation, 4) with the data and compute intensive nature of geospatial processing, new computing paradigms will have to be explored (NRC 2010). In particular, cloud computing seems to be a promising option to handle the complex scalability issues of the future CI (Yang et al., 2011). The key characteristics of cloud computing (on-demand scalability, intra-cloud access to Big Data, and pay-as-you-go model, Yang et al., 2011a) are extremely useful in addressing the requirements of EarthCube.

Thinking and computing in a spatiotemporal fashion will provide an enabling capability for the new geoscience frontier by contributing essential computing architectures, algorithms, and methodologies to construct the CI for solving problems with characteristics of data intensity, computing intensity, spatiotemporal intensity, and concurrent intensity (Yang et al., 2011b). This challenge requires scientists, engineers, and educators from multiple domains to collaborate to solve fundamental problems, e.g., how to model high-resolution phenomena with broad geographic coverage for regional emergency responses, such as tsunami events.

Analyzing geoscience data has been a prominent challenge for decades. The growth of data and the increasing complexity of analysis algorithms have been the most significant problems in this area. Future research on data analysis is likely to center on new analyses capabilities, tools, and platforms in the next decade (Zhang et al., 2003). For example, the NASA Earth Exchange (NEX, 2011) initiative uses HPC to analyze geoscience data. Cloud computing and grid computing are integrated to address geoscience problems, such as climate studies (Climate@Home).

To investigate the connections and interactions across different Earth subsystems, such as the geosphere and biosphere, an interdisciplinary research approach is required. A scientific model workflow to chain independent processes would be very helpful to recruit the needed models for a specific scientific application or task on the fly, ideally, in an automatic fashion. Within a CI with proper metadata and model configurations for existing domain models which have been running separately for different domains, the implementation of interdisciplinary modeling and knowledge sharing would become real.

[19] http://eie.cos.gmu.edu/c/portal/layout?p_l_id=PUB.1.92
[20]http://swp.gmu.edu/climate@home
[21]http://testbed.gmu.edu:9090
[22]http://testbed.gmu.edu

References

1. Donner, R., et al., 2009. Understanding the Earth as a Complex System -recent advances in data analysis and modelling in Earth sciences. *European Physical Journal Special Topics*, 174, 1-9.

2. Goodchild, M., 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69 (4), 211-221.

3. Goodchild, M., Yuan, M., and Cova, T., 2007. Towards a general theory of geographic representation in GIS. *International Journal of Geographic Information Science*, 21, 239-260.

4. Horvitz, E., and Mitchell, T., 2011. From Data to Knowledge to Action: A Global Enabler for the 21st Century. *Computing Community Consortium (CCC)'s white papers*. http://www.cra.org/ccc/docs/init/From_Data_to_Knowledge_to_Action.pdf (accessed on 10/13/2011).

5. Kumar, V., 2007. High performance data mining - application for discovery of patterns in theglobal climate system. In: S. Aluru, M. Parashar, R. Badrinath, and V.K..Prasanna, eds. *Proceedings of the 14th international conference on High performance computing (HiPC'07)*. Berlin, Heidelberg: Springer-Verlag, 4.

6. Nebert, D., and A. Whiteside, 2005. Catalog Services, Version 2, *OGC Implementation Specification*, URL: http://portal.opengis.org/files/?artifact_id5929 (accessed on 10/13/2011).

7. NEX, 2011. *NASA Earth Exchange*, http://www.nasa.gov/centers/ames/news/releases/2010/10-33AR.html.

8. NRC, 2003. *Living on an active earth: Perspectives on earthquake science*. Washington DC: The National Academies Press.

9. NRC, 2009a. *Global environmental health: research gaps and barriers for providing sustainable water, sanitation, and hygiene services: workshop summary*. Washington DC: The National Academies Press.

10. NRC, 2009b. *Informing decisions in a changing climate*. Washington DC: The National Academies Press.

11. NRC, 2010. *The rise of games and high performance computing for modeling and simulation*. Washington DC: The National Academies Press.

12. NRC, 2011. *Tsunami warning and preparedness: an assessment of the U.S. Tsunami Program and the Nation's Preparedness Efforts*. Washington DC: The National Academies Press.

13. NSF, 2009a. *NSF GEO Vision*, http://www.nsf.gov/geo/acgeo/geovision/start.jsp (accessed on 10/13/2011).

14. NSF, 2009b. *NSF-supported research infrastructure: Enabling discovery, innovation and learning*, NSF-09-13, 148 p. http://www.nsf.gov/pubs/2007/nsf0728/index.jsp (accessed on 10/13/2011).

15. NSF, 2011. *Earth Cube Guidance for the Community*, http://www.nsf.gov/pubs/2011/nsf11085/nsf11085.pdf (last date accessed 10/13/2011).

16. OGC, 1998. The OpenGIS Guide: Introduction to Interoperable Geoprocessing.

17. Poore B., 2011. Users as essential contributors to spatial cyberinfrastructures. *Proceedings of National Academy of Sciences of USA*, 108(14), doi: 10.1073/pnas.0907677108.

18. Raskin R. and M. Pan, 2005. Knowledge representation in the semantic web for Earth and environmental terminology (SWEET), *Computers & Geosciences*, 31(9), 1119-1125.

19. Yang, C., Raskin, R., Goodchild, M.F., and Gahegan, M., 2010. Geospatial Cyberinfrastructure: Past, Present and Future, *Computers, Environment, and Urban Systems*, 34(4):264-277.

20. Yang C., Goodchild M., Huang Q., Nebert D., Raskin R., Bambacus M., Xu Y., Fay D., 2011a. Spatial Cloud Computing – How can geospatial sciences use and help to shape cloud computing, *International Journal of Digital Earth*. 4(4), 305-329.

21. Yang C., Wu H., Li Z., Huang Q., Li J., 2011b, Utilizing Spatial Principles to Optimize Distributed Computing for Enabling Physical Science Discoveries, *Proceedings of National Academy of Sciences of USA*, 108(14): 5498-5503.

22. Yang C., Nebert D., Taylor F., 2011c. Establishing a sustainable and cross-boundary geospatial cyberinfrastructure to enable polar research, *Computers & Geosciences*, doi:10.1016/j.cageo.2011.06.009.

23. Zhang C., Zhao T., Li W., Osleeb J., 2010. Towards logic-based geospatial feature discovery and integration using web feature service and geospatial semantic web. *International Journal of Geographical Information Science*, 24(6): 903-923.

24. Zhang P., Huang Y., Shekhar S. and Kumar V., 2003. Correlation Analysis of Spatial Time Series Datasets: A Filter-and-Refine Approach. In: *Proceedings of the 7th Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Berlin, Heidelberg: Springer-Verlag, 532-544.