COVER SHEET FOR PROPOSAL TO THE NATIONAL SCIENCE FOUNDATION

NSF 06-577 09/15/09

FOR CONSIDERATION BY NSF ORGANIZATION UNIT(S) (Indicate the most specific unit known, i.e. program, division, etc.)

BCS - DEL, IIS - SOCIAL-COMPUTATIONAL SYSTEMS

DATE RECEIVED NUMBER OF COPIES DIVISION ASSIGNED FUND CODE DUNS# (Data Universal Numbering System) FILE LOCATION
09/15/2009 2 04040000 BCS 7719 038633251 09/15/2009 8:27pm

EMPLOYER IDENTIFICATION NUMBER (EIN) OR TAXPAYER IDENTIFICATION NUMBER (TIN)
141368361

SHOW PREVIOUS AWARD NO. IF THIS IS A RENEWAL

IS THIS PROPOSAL BEING SUBMITTED TO ANOTHER FEDERAL AGENCY?
NO

NAME OF ORGANIZATION TO WHICH AWARD SHOULD BE MADE
SUNY at Buffalo

ADDRESS OF Awardee ORGANIZATION, INCLUDING 9 DIGIT ZIP CODE
501 Capen Hall
Buffalo, NY 14260-0000

Awardee ORGANIZATION CODE (IF KNOWN)
0028373000

NAME OF PERFORMING ORGANIZATION, IF DIFFERENT FROM ABOVE

ADDRESS OF PERFORMING ORGANIZATION, IF DIFFERENT, INCLUDING 9 DIGIT ZIP CODE

PERFORMING ORGANIZATION CODE (IF KNOWN)

IS Awardee ORGANIZATION (Check All That Apply)

SMALL BUSINESS

MINORITY BUSINESS

WOMAN-OWNED BUSINESS

IF THIS IS A PRELIMINARY PROPOSAL THEN CHECK HERE

TITLE OF PROPOSED PROJECT
DEL/SoCS: Extending Preservation Archives through Social Networking

REQUESTED AMOUNT
$ 449,594

PROPOSED DURATION (1-60 MONTHS)
36 months

REQUESTED STARTING DATE
06/01/10

SHOW RELATED PRELIMINARY PROPOSAL NO. IF APPLICABLE

CHECK APPROPRIATE BOX(ES) IF THIS PROPOSAL INCLUDES ANY OF THE ITEMS LISTED BELOW

BEGINNING INVESTIGATOR (GPG I.G.2)

DISCLOSURE OF LOBBYING ACTIVITIES (GPG II.C.1.e)

PROPRIETARY & PRIVILEGED INFORMATION (GPG II.D.1.d)

HISTORIC PLACES (GPG II.C.2.j)

EAGER* (GPG II.D.2) RAPID** (GPG II.D.1)

VERTEBRATE ANIMALS (GPG II.D.6) IACUC App. Date

PHS Animal Welfare Assurance Number

HumAn Subjects (GPG II.D.7) Human Subjects Assurance Number FWA00008824

Exemption Subsection or IRB App. Date 05/21/09

INTERNATIONAL COOPERATIVE ACTIVITIES: COUNTRY/COUNTRIES INVOLVED (GPG II.C.2.j)

HIGH RESOLUTION GRAPHICS/OTHER GRAPHICS WHERE EXACT COLOR REPRESENTATION IS REQUIRED FOR PROPER INTERPRETATION (GPG I.G.1)

P/I/PD DEPARTMENT
CSE

P/I/PD POSTAL ADDRESS
201 Bell Hall
Buffalo, NY 142602000

United States

P/I/PD FAX NUMBER
716-645-3464

NAMES (TYPED)

Michail Petropoulos PhD 2006 716-645-4733 mpetropo@buffalo.edu

CO-P/I/PD

Jeffrey C Good PhD 2003 716-645-2177 jcgood@buffalo.edu

PI/PD NAME

Michail Petropoulos

CO-P/I/PD

Jeffrey C Good

Electronic Signature
**CERTIFICATION PAGE**

**Certification for Authorized Organizational Representative or Individual Applicant:**

By signing and submitting this proposal, the Authorized Organizational Representative or Individual Applicant is: (1) certifying that statements made herein are true and complete to the best of his/her knowledge; and (2) agreeing to accept the obligation to comply with NSF award terms and conditions if an award is made as a result of this application. Further, the applicant is hereby providing certifications regarding debarment and suspension, drug-free workplace, and lobbying activities (see below), nondiscrimination, and flood hazard insurance (when applicable) as set forth in the NSF Proposal & Award Policies & Procedures Guide, Part I: the Grant Proposal Guide (GPG) (NSF 09-29). Willful provision of false information in this application and its supporting documents or in reports required under an ensuing award is a criminal offense (U. S. Code, Title 18, Section 1001).

**Conflict of Interest Certification**

In addition, if the applicant institution employs more than fifty persons, by electronically signing the NSF Proposal Cover Sheet, the Authorized Organizational Representative of the applicant institution is certifying that the institution has implemented a written and enforced conflict of interest policy that is consistent with the provisions of the NSF Proposal & Award Policies & Procedures Guide, Part II, Award & Administration Guide (AAG) Chapter IV.A; that to the best of his/her knowledge, all financial disclosures required by that conflict of interest policy have been made; and that all identified conflicts of interest will have been satisfactorily managed, reduced or eliminated prior to the institution’s expenditure of any funds under the award, in accordance with the institution’s conflict of interest policy. Conflicts which cannot be satisfactorily managed, reduced or eliminated must be disclosed to NSF.

**Drug Free Work Place Certification**

By electronically signing the NSF Proposal Cover Sheet, the Authorized Organizational Representative or Individual Applicant is providing the Drug Free Work Place Certification contained in Exhibit II-3 of the Grant Proposal Guide.

**Debarment and Suspension Certification**

<table>
<thead>
<tr>
<th>Question</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is the organization or its principals presently debarred, suspended, proposed for debarment, declared ineligible, or voluntarily excluded from covered transactions by any Federal department or agency?</td>
<td>☐</td>
<td>☑</td>
</tr>
</tbody>
</table>

By electronically signing the NSF Proposal Cover Sheet, the Authorized Organizational Representative or Individual Applicant is providing the Debarment and Suspension Certification contained in Exhibit II-4 of the Grant Proposal Guide.

**Certification Regarding Lobbying**

The following certification is required for an award of a Federal contract, grant, or cooperative agreement exceeding $100,000 and for an award of a Federal loan or a commitment providing for the United States to insure or guarantee a loan exceeding $150,000.

**Certification for Contracts, Grants, Loans and Cooperative Agreements**

The undersigned certifies, to the best of his or her knowledge and belief, that:

1. No federal appropriated funds have been paid or will be paid, by or on behalf of the undersigned, to any person for influencing or attempting to influence an officer or employee of any agency, a Member of Congress, an officer or employee of Congress, or an employee of a Member Congress in connection with the awarding of any federal contract, the making of any Federal grant, the making of any Federal loan, the entering into of any cooperative agreement, and the extension, continuation, renewal, amendment, or modification of any Federal contract, grant, loan, or cooperative agreement.

2. If any funds other than Federal appropriated funds have been paid or will be paid to any person for influencing or attempting to influence an officer or employee of any agency, a Member of Congress, an officer or employee of Congress, or an employee of a Member of Congress in connection with this Federal contract, grant, loan, or cooperative agreement, the undersigned shall complete and submit Standard Form-LLL, “Disclosure of Lobbying Activities,” in accordance with its instructions.

3. The undersigned shall require that the language of this certification be included in the award documents for all subawards at all tiers including subcontracts, subgrants, and contracts under grants, loans, and cooperative agreements and that all subrecipients shall certify and disclose accordingly.

This certification is a material representation of fact upon which reliance was placed when this transaction was made or entered into. Submission of this certification is a prerequisite for making or entering into this transaction imposed by section 1352, Title 31, U. S. Code. Any person who fails to file the required certification shall be subject to a civil penalty of not less than $10,000 and not more than $100,000 for each such failure.

**Certification Regarding Nondiscrimination**

By electronically signing the NSF Proposal Cover Sheet, the Authorized Organizational Representative is providing the Certification Regarding Nondiscrimination contained in Exhibit II-6 of the Grant Proposal Guide.

**Certification Regarding Flood Hazard Insurance**

Two sections of the National Flood Insurance Act of 1968 (42 USC §4012a and §4106) bar Federal agencies from giving financial assistance for acquisition or construction purposes in any area identified by the Federal Emergency Management Agency (FEMA) as having special flood hazards unless the:

1. Community in which that area is located participates in the national flood insurance program; and
2. Building (and any related equipment) is covered by flood insurance.

By electronically signing the NSF Proposal Cover Sheet, the Authorized Organizational Representative or Individual Applicant located in FEMA-designated special flood hazard areas is certifying that adequate flood insurance has been or will be obtained in the following situations:

1. For NSF grants for the construction of a building or facility, regardless of the dollar amount of the grant; and
2. For other NSF Grants when more than $25,000 has been budgeted in the proposal for repair, alteration or improvement (construction) of a building or facility.

<table>
<thead>
<tr>
<th>Name</th>
<th>Signature</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jeffrey S Schieder</td>
<td>Electronic Signature</td>
<td>Sep 15 2009 4:49PM</td>
</tr>
</tbody>
</table>

**Telephone Number** 716-645-5000  **Electronic Mail Address** Jeffrey.Schieder@buffalo.edu  **Fax Number** 716-645-2760

* EAGER - Early-concept Grants for Exploratory Research
** RAPID - Grants for Rapid Response Research
Intellectual Merit: In recent decades, the field of linguistics has become increasingly concerned with the development of infrastructure to support the long-term maintenance of language resources, especially resources documenting endangered languages. At the same time, the field has also recognized the importance of ensuring that these resources are accessible to as wide an audience as possible, both academic and non-academic and, in particular, to communities whose languages are represented in those resources.

The Web offers numerous examples (e.g., Facebook) of the possibilities for user interaction with heterogeneous data sets of the sort likely to be contained in a language archive (e.g., digitized recordings, transcriptions, translations, etc.). However, communities based around an archive are not necessarily based around shared general interests—the canonical case for online web communities—but rather shared use of resources in the archive. Therefore, existing models need to be adapted to allow for resource-based, rather than interest-based organization. Second, material on the Web is generally construed as ephemeral, without a commitment to longevity, which is one of the defining characteristics of an archive.

The primary goal of the proposed project is to conduct research on how to form a successful pairing of the core preservation functions of a digital archive with modern community building platforms, social network analysis, and semantic web techniques so that the archive’s collections can be extended, enriched and effectively mobilized for research through community use.

Community-Based Archive Extension: If an archive is to exploit web-based methods to facilitate resource mobilization, what is required is the development of a system that builds on existing data dissemination methods in ways that make them compatible with its mission of long-term preservation. A particular area of interest is the extent to which web-based community tools could allow users of the archive to create content which itself might make a valuable contribution to the archive. While much user-generated content may be too specific or personal to be useful in the archive, in some cases users may produce content of general, long-term value (e.g., a translation of a recording). In principle, such material should then be incorporated into the archive itself and given the same commitment to permanence as other materials in the collection.

The core problem here consists of identifying community resources of high enough “quality” to become archival resources, and subsequently promoting them into archival resources in a consistent and meaningful manner. We will expand on preliminary research to develop a semi-automatic solution to the problem, where initially a set of high quality community resources are automatically identified and recommended to an archivist for promotion. In the second step, the archivist manually examines the recommendations, decides which ones to promote and establishes relationships between the promoted and the existing resources.

Open Preservation and Extension Platform: The proposed research will be deployed as part of an open preservation platform, which will be designed to facilitate rapid publishing and promotion of resources in preservation archives. The platform to be developed will be tested in the context of the developing Northeastern North American Indigenous Languages Archive (NNAILA). NNAILA’s goals are to preserve recordings of indigenous languages of Northeastern America and to make the data in those recordings accessible to the academic community and to speaker communities. NNAILA has digitized over 100 hours of recordings of Onondaga and developed a prototype semantic web portal to allow users to access materials in those recordings. Work already done on this portal will form the foundation for this project.

Broader Impacts: The proposed research will develop and distribute infrastructure enabling the controlled sharing of data on the Web. The modular architecture of the proposed platform will also be exploited in teaching, tying together various courses from the standard computer science and linguistics curriculum. In addition, since the target resource domain for this project involves materials from indigenous languages, their increased accessibility will allow both scholars and community members to use the materials employing powerful social network methods previously unavailable to them.
<table>
<thead>
<tr>
<th>Section</th>
<th>Total No. of Pages</th>
<th>Page No.* (Optional)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cover Sheet for Proposal to the National Science Foundation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Project Summary (not to exceed 1 page)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Table of Contents</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Project Description (Including Results from Prior NSF Support) (not to exceed 15 pages) <em>(Exceed only if allowed by a specific program announcement/solicitation or if approved in advance by the appropriate NSF Assistant Director or designee)</em></td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>References Cited</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Biographical Sketches (Not to exceed 2 pages each)</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Budget (Plus up to 3 pages of budget justification)</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Current and Pending Support</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Facilities, Equipment and Other Resources</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Special Information/Supplementary Documentation</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Appendix (List below.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>(Include only if allowed by a specific program announcement/solicitation or if approved in advance by the appropriate NSF Assistant Director or designee)</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appendix Items:</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Proposers may select any numbering mechanism for the proposal. The entire proposal however, must be paginated. Complete both columns only if the proposal is numbered consecutively.*
1 Introduction

In recent decades, the field of linguistics has become increasingly concerned with the development of infrastructure to support the long-term maintenance of language resources, especially resources documenting endangered languages [11, 12]. At the same time, the field has also recognized the social importance of ensuring that these resources are accessible to as wide an audience as possible, in particular, to communities whose languages are represented in those resources, many of which can put them to direct use in language maintenance programs if they can be transformed to appropriate formats [42, 25]. This line of thinking is particularly noticeable in emerging models of linguistic work which adopt modes of research involving substantial collaboration between researchers and communities, and it is clear that computational techniques can have a substantial role to play in facilitating such collaboration [41, 19, 63, 51, 16].

The Web offers numerous examples (e.g., Facebook) of the possibilities for user interaction with heterogeneous data sets of the sort likely to be contained in a language archive (e.g., digitized recordings, transcriptions, translations, etc.). In practice, however, such models cannot be directly exploited for use by an archive for two reasons. First, communities based around an archive are not necessarily based around shared general interests—the canonical case for online web communities—but rather shared use of resources in the archive. Therefore, existing models need to be adapted to allow for resource-based, rather than interest-based organization. Second, material on the Web is generally construed as ephemeral, without the commitment to longevity that is one of the defining characteristics of an archive [30].

The primary goal of the proposed project is to conduct research on how to form a successful pairing of the core preservation functions of a digital archive with modern community building platforms, social network analysis, and semantic web techniques so that the archive’s collections can be extended, enriched and effectively mobilized for research through community use. It involves novel collaboration between a linguist with expertise in digital linguistics and a computer scientist with an expertise in databases who will collaborate with a Native American community, and will produce not only scientific results but will also (i) ensure that substantial linguistic materials presently recorded on perishable media will be preserved in the long-term, (ii) create an online portal to allow members of the collaborating community and appropriate researchers to access those resources and create new ones, and (iii) produce tools allowing other projects to create portals with similar aims and structures. Because of the project’s focus on building a system to facilitate use of endangered language resources by communities of interest formed around these resources, we are submitting this proposal jointly to the NSF’s Documenting Endangered Language (DEL) and Social-Computational Systems (SoCS) programs.

Section 2 introduces the project investigators. Section 3 describes the institutional context in which this research will take place and gives relevant background on digital archiving of language resources. Section 4 overviews the structure of the computational system which is being developed by this project, and Section 5 discusses the technical work that has been done to date on its development and work to be done in the near future. Section 6 describes research to be conducted to ensure that the computational system will ultimately be able to serve the widest range of linguistic communities possible. Section 7 outlines the research and evaluation plan and summarizes the project deliverables. Section 8 discusses the project’s broader impacts.

2 Collaborating Investigators

Jeff Good is Assistant Professor in the Department of Linguistics at the University at Buffalo. He has a long-standing interest in digital archiving and extensive experience in the development and use of best-practice standards and methodologies in documentary and descriptive linguistics, the design and implementation of linguistic databases, as well as in traditional descriptive linguistics.
He has held an NEH Documenting Endangered Languages (DEL) Fellowship (FN #5000506) to do documentary and descriptive work on the Western Beboid languages of Cameroon and was funded by the NEH (RZ-50817-07) to do salvage documentation on the moribund languages of the Furu Awa region of Cameroon. Both projects have resulted in the collection of the first archived materials on these languages, and initial results of the work on have been reported [22, 27, 33]. He is continuing this line of work in a currently funded DEL project Towards an areal grammar of Lower Fungom (NSF BCS-0853981).

In addition, he serves as co-PI for the NSF-funded lexicon interoperability project, Lexicon Enhancement via the GOLD Ontology (LEGOD; BCS-0753321), and Co-PI of the AnthropDataDPA project (NSF BCS-0823404), which has the goal of developing a comprehensive data preservation and access plan for anthropological data of all kinds. He also served as PI on the Pangloss project (NSF BCS-0715246), which has produced a feasibility study on the use of OpenOffice.org as a platform for linguistic tool development. Since arriving at the University at Buffalo in 2007, Good has worked to establish a digital archive of language materials for languages traditionally spoken in the northeastern North America, NNAILA (see Section 3.1). This work emanated naturally from his earlier work on the Rosetta Project (funded by NSF DUE-0333727), for which he serves as technical adviser [26]. He is a member of the board of the Digital Endangered Languages and Musics Network (DELAMAN) and chair of the Linguistic Society of America’s Technology Advisory Committee. During the academic year 2009–2010 he is serving as interim director of the University at Buffalo’s Digital Humanities Initiative.

Michalis Petropoulos is Assistant Professor in the Computer Science and Engineering Department at the University at Buffalo. His primary research area is in the intersection of Web and databases. He is particularly interested in reducing the cost of data publishing and semantic integration, and enable the exploration and maintenance of large data sets through the use of web-based interactive visual interfaces, a need that has been well documented by the database community [3, 8].

His research is practically motivated and has resulted in five interactive systems that are publicly available. In data integration, he developed the CLIDE system [52, 53] for interactively querying web-service accessed data sources, which was well received by the database community earning a Best Paper Award Honorable Mention at the ACM International Conference on Management of Data (SIGMOD), in 2006. In data exploration, his research group developed the BioNav system [31, 32] for effective navigation of biomedical databases, such as MEDLINE, using concept hierarchies and ontologies, such MeSH and GO. He has also built the QUROSED system [50, 54] for querying and reporting XML/semi-structured data on the Web, which is described in a patent assigned to Oracle Corp., and the Protoplasm system [9] while at Microsoft Research. Finally, his group is developing the Northeastern North American Indigenous Languages Archive (NNAILA) portal [44] described in this proposal.

He has been publishing in both web and database systems venues, such as the ACM Transactions on Database Systems, the ACM Transactions on Internet Technology, the ACM Conference on Management of Data (SIGMOD), and the International World Wide Web (WWW) conference. In 2009, he co-chaired the Twelfth International Workshop on Web and Databases (WebDB), one of the first forums dedicated to researchers, theoreticians, and practitioners working on the intersection of data management and the Web, held together with the ACM SIGMOD/PODS conference. He is currently serving in the program committee of the International World Wide Web Conference (WWW 2010) and has served as the tutorials chair for the International Conference on Web Engineering (ICWE), in 2007.

3 Digital Language Archiving
3.1 The Northeastern North America Indigenous Languages Archive

The Northeastern North American Indigenous Languages Archive (NNAILA) is a new digital language archive housed at the University at Buffalo (UB) within the University Library system. NNAILA is currently
in its pilot phase, with a focus on digitizing materials from Onondaga [ono], an Iroquoian language spoken at Onondaga Indian Nation near Syracuse, New York and at the Six Nations Indian Reserve near Brantford, Ontario by less than 100 fluent speakers [40]. The archive’s primary goals are to preserve recordings of indigenous languages of Northeastern North America and to make those recordings accessible to researchers and the indigenous communities whose languages are represented in the archive’s collections.

UB is an ideal home for an archive with such a focus due both to the strength of the linguistics department with respect to language documentation and description and the university’s central location within northeastern North America. While there are a number of existing archives with materials on the languages of northeastern North America (e.g. the American Philosophical Society), there is, as yet, no dedicated digital archive for these languages, even though they are situated in a part of the world where not only scholars, but also community members, are in a position to make use of the potential that digital resources offer for internet access. The archive aims to fill this gap by offering digitization services for analog materials, with an initial focus on digitization of audio recordings on perishable media like cassettes, and serving as a central repository for preservation and access of resources on the languages of this part of the world.

In order to facilitate access to the archive’s resources by community members, especially language teachers and their students, NNAILA is developing a web portal which will allow users to make use of the archive’s collections [44] (see Section 4). The portal has been applied primarily to the archive’s Onondaga materials and is being tested in the Onondaga language program. As its development stabilizes, NNAILA will attempt to apply it more broadly (see Section 6).

NNAILA has received pilot funding totaling more than $50,000 from grants from a number of sources internal to UB. These grants have funded digitization and tool development for the project to this point. Almost 200 hours of Onondaga material, collected primarily by Dr. Hanni Woodbury and Dr. Percy Abrams (who is also a member of the Onondaga Nation and head of their language program) have been digitized and are being archived. In the fall of 2008, the UB Department of Linguistics set aside office space for a dedicated digitization station for the archive. In the summer of 2009, the NNAILA project team made a visit to the Onondaga Nation to do a demonstration of the developing portal to a group of community members training to become language teachers and input from that visit has guided the specifics of some of the work proposed here. We have also received extensive and valuable input from one of NNAILA’s main depositors during the pilot phase, Dr. Hanni Woodbury.

3.2 Broader Developments
The development of NNAILA is occurring as part of a broader trend within linguistics to establish digital preservation archives. The real-world impetus behind this trend is two-fold. On the one hand, the accelerating pace of language endangerment and extinction [34] has triggered an increase in efforts to document the world’s languages, as evidenced by such initiatives as the Dokumentation Bedrohter Sprachen (DoBeS) project [17], the Hans Rausing Endangered Languages Program [58], and the NSF/NEH Documenting Endangered Languages (DEL) program. A core aim of these efforts is to send teams of researchers to locations throughout the world to collect data from a wide variety of endangered languages. This data is expensive to collect. Accordingly, initiatives like these have requirements that it be deposited in a preservation archive.3

1 Due to sensitivities, it is not possible for us to give access to the materials hosted on the NNAILA portal to anonymous reviewers, but we have created a special reviewers’ area, accessible at http://nnaila.org/reviewer with username reviewer and password nsf, where the current functionality is demonstrated with an Onondaga recording that was made public by the American Philosophical Society.

2 Specifically, 127 hours of recordings from Hanni Woodbury’s collections, fifty-seven hours from Percy Abrams’ collections, and four hours from Dr. Karin Michelson’s collections have been digitized.

3 The guidelines for the DEL program contain explicit instructions for proposers to, “Discuss plans for archiving recordings, field notes, and processed documentary materials in a stable environment.” (See http://www.nsf.gov/pubs/2006/nsf06577/nsf06577.htm.)
At the same time, the increasing pace of language endangerment has also made analog recordings of endangered languages more valuable insofar as it is likely that many of them will soon become irreplaceable. This has had two prominent effects. First, extensive resources have been put into educating the linguistics community about appropriate methods for migrating legacy language resources to archival digital formats, most prominently advocated in the context of the NSF-funded Electronic Metastructure for Endangered Languages Data (E-MELD) project (NSF BCS-0094934; [20, 12]). Second, in recent years several digital archives have been founded which have, among their primary goals, to assist in the digitization of analog recordings to digital formats for the purpose of long-term preservation.4

These developments within linguistics not only mean that NNAILA is operating in a generally favorable research climate but also that the tools to be produced via this project have the potential to be put to immediate use elsewhere, especially given the additional trend within linguistics towards ensuring that the products of linguistic research also serve community interests to the extent possible, as discussed in Section 1.

3.3 Resource Sensitivities
When working on a project involving a diverse set of stakeholders, from researchers, to a small group of (usually elder) fluent speakers, to community members who do not speak the language but view it as an important part of their heritage, there will inevitably be a complex array of sensitivities to be monitored, especially in cases when a researcher is a member of a larger community that has been marginalizing indigenous communities for centuries [62]. Accordingly, the initial stages of NNAILA have proceeded with the involvement a relatively small and targeted group of individuals in order to allow us to assess sensitivities in a controlled fashion.

On the community side, our principal liaison is Dr. Percy Abrams an Onondaga linguist who runs the Onondaga Nation language program and has research interests in using his linguistic training to facilitate language learning [1]. His participation gives us a point of contact to the community who has a clear understanding both of the culture of academic linguistics and how the work of linguists could positively (or not) impact the Onondaga community. Dr. Abrams has additionally put the project in touch with members of the Onondaga Nation training to become language teachers themselves, and they have served as an initial test audience for the tools we have developed. On the researcher side, NNAILA has benefitted from the continuing advice of Dr. Karin Michelson, a specialist of Oneida [one], another Iroquoian language, who is based at the University at Buffalo, and Dr. Hanni Woodbury, a specialist of Onondaga who has allowed some of her materials to be used in the initial stages of the project.

NNAILA has made, and will continue to make, efforts to understand and address the wishes of all stakeholders regarding access to and uses of its materials. In many cases, this will mean restricting access to the materials as well as to any user-created contributions to the materials. While such restrictions may mean that some of the information created as a result of this research cannot be made fully public, the bulk of the research results can be separated from the specific content of the portal, thus allowing this project to respect resource sensitivities and still produce significant results.

4 Community-Based Archive Extension
If an archive is to exploit web-based methods to facilitate use of resources by diverse communities, what is required is the development of a system that (i) builds on existing data dissemination methods in ways that make them compatible with its mission of long-term preservation, (ii) is sensitive to linguistic and user interaction concerns, and (iii) addresses digital infrastructure issues. A particular area of interest in the present project is the extent to which web-based community tools could allow users of the archive to create

4 Examples include the Archive of the Indigenous Languages of Latin American [57] and the Pacific And Regional Archive for Digital Sources in Endangered Cultures [48].
content which itself might make a valuable contribution to the archive. While much user-generated content may be too specific or personal to be useful in the archive, (e.g., a comment stating that a user likes or does not like a given resource), in some cases users may produce content of general, long-term value (e.g., a translation of a recording). In principle, such material should then be incorporated into the archive itself and given the same commitment to permanence as other materials in the collection.

The overall vision coming out of these observations is the NNAILA hybrid system architecture depicted in Figure 1. In addition to having a core Archive component, residing in long-term preservation storage (in a so-called “dark” archive), NNAILA enables publishing the resources in the archive online for use by a Community Portal in a controlled manner. A researcher or a language teacher with appropriate access permissions chooses a subset of the archive resources to be published on a Community Portal using the Collection Building Tools, which transform the selected resources in a format compatible with the portal. Once published, the archive resources can be accessed by diverse groups of users, such as speaker communities, students of a language class, and academic researchers (assuming they have the relevant permissions). A valid user can then contribute to the archive resources in the collection by annotating them with their own content. Note that within the Community Portal in Figure 1, the archive resources (solid circles) and the user contributions (hollow circles) coexist within the same storage space and are semantically connected to each other. (Not explicitly depicted is the fact that archive resources published in the portal maintain links to their surrogates in the “dark” archive.)

In time, the community users would submit a structured set of contributions to published archive resources, some of which would contain content of high enough “quality” to be promoted to archival status and migrated into the archive (see Section 4.2 for further discussion of the issue of “quality”). This semi-automatic promotion process is shown in Figure 1. NNAILA monitors the activity at the Community Portal, detects user contributions of potentially high quality and proposes them to the Archivist. Subsequently, and through the use of the Promotion Tools, the Archivist promotes the content of one or more user contributions to a new archive resource, which is migrated into the Archive component.

At present, no preservation archive that we are aware of supports such functionality. A previous version of the archive of the Rosetta Project [59] came close by associating discussion forums with individual

---

5 At present, the linguistic personnel involved in the development of NNAILA are also serving as its archivists. However, as the archive develops, it will seek out the resources needed to hire a dedicated archivist.
languages using a semantic web architecture [26], but it had neither generalized community building tools nor a mechanism through which user-generated content could be selectively migrated into the archive. The functionality schematized in Figure 1 would, therefore, be novel and clearly desirable from the point of view of the archive itself and its users.

We discuss next the two processes, and the components involved, supported by the NNAILA system architecture in Figure 1, namely publishing archive resources and promoting community user contributions. For each one, we elaborate on the current state of development achieved by prior funding and list the open issues we plan to work on.

4.1 Publishing Archive Resources

Currently, NNAILA uses semantic web technologies and techniques [7, 36, 24, 38] to model and store archive resources, such as audio recordings, images, transcriptions, and translations, and to express the relationships among them. The NNAILA Community Portal employs an existing content management and social networking platform [15] to publish archive resources on the Web and enable community users to contribute to them by submitting their own content in various forms, such as text, images, audio files, etc.

Figure 2 shows an audio archive resource of a recording of an Onondaga story entitled **HW06 Tall Corn**, as published by the current implementation of the NNAILA Community Portal [44]. Apart from providing the ability to play the audio resource, the portal also displays related metadata, such as creation date, dialect and the speaker's name. Also shown on the top right corner are archive resources related to the one currently being viewed. For example, archive resource **LG01 Locusts** is related because it discusses a Similar Topic As the current resource, and **NC02 The Longlegged Deer** was recorded during the Same Session As the current resource.

After listening to the audio archive resource, community users submit their contributions. The NNAILA Community Portal enables them to contribute content related to the whole audio resource or part of it. For example, the first contribution shown in Figure 2 is submitted by community user **ck0** and refers to the initial 26 seconds of the audio resource’s total play time of 2 minutes and 47 seconds. Other community users either reply to submitted contributions or add their own using the corresponding hyperlinks. They can also express the helpfulness of a contribution through the opinion mechanism incorporated and displayed within every user contribution. Moreover, community users can tag archive resources, as is the case with tag **corn** in Figure 2.

The Collection Building Tools component, also involved in the publishing process shown in Figure 1, is not as developed as the Community Portal. A researcher or a teacher can only choose whether to publish an archive resource on the Community Portal or not. Also, the only two user groups currently supported by the Community Portal are the administrators, such as a researcher, teacher or the Archivist, and the community users, such as students, speaker community members, etc. In Section 5, we discuss in detail which of the above NNAILA Community Portal and Collection Building Tools functionality we developed in-house, such as support for audio resources and semantic relationships, and what was provided to us by existing platforms [15], such as tracking the number of times a resource has been viewed.

**Open Issues** What follows is the list of issues we plan to work on and are related to the process of publishing archive resources. They will be developed as part of either the Collection Building Tools or the Community Portal components. While some issues are more of an engineering nature, others require novel research to be conducted before implementation can be realized.

---

6 While the display configuration in Figure 2 only shows a short user name, more information about each user can be found by clicking on the user name. Consultation with researchers and community members clearly indicated that a lack of anonymous contributions was a desirable feature of the system.
Figure 2: An Audio Archive Resource and User Contributions on the NNAILA Community Portal
• Support for *multiple collections* of published archive resources. For example, two teachers should be able to create two separate resource collections that will be accessed by their respective students. Consequently, the contributions made to these resources should also be within the context of a given collection.

• Specify and develop a *resource-oriented* role-based access control mechanism (RBAC [23]). A researcher or a teacher should be able to specify which resources in a collection are accessible by which user groups.

• Design and implement *audio snipping*. Some of the recordings in the NNAILA archive are very long (over 40 minutes) or may capture more than one “logical” resource (e.g., multiple stories). Therefore, it may be desirable to create new audio resources that encompass only a portion of a given recording, such as a single story. This functionality will provide a web-based interface to researchers and teachers creating collections, and will automatically record the provenance of the new resources. The timestamps associated with these audio “snippets” will be treated as promotable data (see Section 4.2) since the partitions they define may help increase accessibility to the content of the recordings.

• Support *offline batch upload*. The NNAILA Archive, residing in long-term preservation storage, consists primarily of several hundred digital audio files and associated metadata. This offline tool will automate the process of converting new recordings in the archive to the web-friendly MP3 format, create the corresponding audio resources in the Community Portal and publish their associated metadata.

• Provide an *archive correction* mechanism. Community users should be able to report errors and suggest corrections for archive resources and their metadata, which subsequently will be examined and executed by the Archivist. (The same basic functionality required for archive correction will also be adapted to allow researchers to submit updates to the metadata of the resources they have deposited.)

### 4.2 Promoting Community User Contributions

The proposed process of promoting community user contributions to archive resources, shown in Figure 1, is semi-automatic and is comprised of two primary steps. First, NNAILA identifies community user contributions of high enough “quality”, within the myriad of contributions submitted at the Community Portal, and proposes them to the Archivist for promotion. Subsequently, the Archivist examines the proposed user contributions and manually executes the promotion to extend the archive in a consistent and meaningful manner. The Promotion Tools, also shown in Figure 1, consist of a search component that performs the first step of the process and a graphical interface that assists the Archivist in performing the second step. In the

![Figure 3: Promoting Community Resources to Archive Resources](image)
following, we discuss the second step of the process first because it will help us build an intuition of what constitutes a “quality” user contribution, which is important for the first step of the process.

**Archive Extension** Under the hood, the NNAILA Community Portal models user contributions as *community resources* and distinguishes them from archive resources while both are stored within a single semantic web graph [36]. NNAILA captures their association though using a general *contribution* relationship type. As an example, Figure 3a displays the semantic web graph that partially models the archive resource and the user contributions/community resources displayed on Figure 2. The archive resource is denoted by the solid circle and is labeled with its title and type (audio), the community resources are denoted by the hollow circles, and the *contribution* relationship type labels the edges connecting them.\(^7\)

Assuming the Archivist has chosen to promote community resources 2-text and 3-text of Figure 3a, corresponding to the second and third contributions on Figure 2, they extend the semantic web graph to reflect the promotion. Executing a promotion involves the following tasks resulting into the extended semantic web graph in Figure 3b:

- The Archivist first curates the chosen community resources by merging and editing their content. Note that a set of community resources can be promoted into a single archive resource. In the case of community resources 2-text and 3-text, the Archivist can merge the two spellings of the word “sun” submitted by community users on Figure 2.
- Following this, a new archive resource is created to accommodate the new content (6-text on Figure 3b).
- The Archivist establishes relationships between the newly created archive resources and existing ones. In Figure 3b, 6-text is linked to archive resource 1-audio via a *transcription* relationship.
- Behind the scenes, the NNAILA Promotion Tools record *promotion provenance* information by automatically establishing *derivedFrom* relationships, as shown in Figure 3b, between the new archive resource 6-text and the promoted community resources 2-text and 3-text. This promotion provenance information will be used to identify quality community resources as described next.

The current implementation of the NNAILA Promotion Tools component does not support the archive extension steps described above, apart from establishing relationships between existing archive resources. We describe in detail what will work on in the open issues list at the end of this section.

**Identifying Quality Community Resources** As the number of user contributions submitted on the NNAILA Community Portal grows, the Archivist will not be able to locate quality community resources by manual inspection. Instead, the NNAILA Promotion Tools should be able to identify quality community resources automatically and propose them to the Archivist for promotion. Providing such functionality might initially seem a hard problem to tackle. For example, how do we identify that the first user contribution in Figure 2 is of low quality, since it does not provide a correct transcription, while the second and third are of high quality? Preferring the user contributions submitted last is not an easy solution. The last user contribution in Figure 2, for example, is completely irrelevant to the other ones. There are arbitrarily many definitions of quality based on which community resources can be ranked [6] and none of them provide any guarantee. On the other hand though, the *social interaction* taking place on the Community Portal and the *promotion provenance* recorded by the Promotion Tools, we believe, points us in promising directions.

In the context of NNAILA, we capture social interaction of a given user as a three-part activity: (i) the resources (and their metadata) for which he or she submits contributions, (ii) the users whose contributions he or she replied to, and (iii) the opinion that other users have of her contributions. The last part of the social interaction is subject to the opinion mechanism employed by NNAILA as discussed in the beginning.

\(^7\) Not shown are the metadata information of the audio archive resource and the thread relationships among community resources evident by the indentation on Figure 2.
of Section 4.8 Promotion provenance information [13, 14, 61] for a given user determines which of their contributions was promoted to archive resources, as well as which other user contributions participated in the promotion.

Having the social interaction and promotion provenance information available, we will develop algorithms intended to identify not only quality community resources but also reliable community users. Existing work on social network analysis [2, 10, 39, 49, 4] will be used as a starting point and extended to identify reliable users predicated upon archive resource metadata. For example, a community user might be reliable for the Onondaga language only and not for others.

Another promising area that we will investigate is that of collaborative filtering (or social filtering) [29, 37, 47, 60]. In principle, collaborative filtering techniques generate item recommendations for a target user based on opinions and ratings of items made by similar users in the community, that is, users sharing similar or highly correlated ratings histories with the target user. In the context of NNAILA, items are the contributions submitted by community users, and the target user is the Archivist. Collaborative filtering cannot be directly applied to NNAILA since the Archivist does not rate any user contributions. Moreover, these techniques target a user within a group with a certain ratings pattern. In NNAILA, the user contributions proposed to the Archivist must be chosen from all such user groups in the community. The direction we plan to pursue is the development of collaborative filtering techniques based on the ratings pattern of the community resources the Archivist has promoted in the past.

Open Issues The following list specifies the core issues we will work on in the area of extending the archive and identifying quality contributions during this project.

- Build a web-based promotion interface to assist the Archivist in performing the promotion tasks outlined above. Existing interfaces [46, 35, 56] are either not web-based or too low level (editing the semantic web graph directly) for practical use.
- Provide interfaces that enable user contributions to promoted resources of “new” types, namely, text and image. Community users should be able to submit a contribution referring to a specific part of a long text resource [55] or a specific region or person on a large image.
- Formally define what constitutes social interaction and promotion provenance information, design corresponding data structures, and devise efficient methods to capture, store and query such information.
- Design and implement identification algorithms for quality contributions and reliable community users based on the information captured above.
- Experiment with alternative opinion mechanisms and find out how they affect the outcome of the identification algorithms, as well as the social interaction of the users on the Community Portal.
- Investigate how to support rich semantic relationships among archive resources, such as one-to-many (ancestor/descendants) and many-to-many relationships amongst speakers, and build a web interface to capture and display such complex relationships.
- Engineer a process to propagate semantic web graph extensions to the Archive in long-term preservation storage (i.e., the “dark” archive).

5 Open Preservation and Extension Platform
The proposed research will result into the development of an Open Preservation and Extension Platform, which is designed to facilitate rapid publishing and promotion of resources in preservation archives while providing clearly defined APIs that will allow third parties to plug in their own content and functionality.

8 In order to address concerns raised by researchers and community members, information about users’ opinions of a given contribution is recorded but not made visible on the site. This limits the possibilities for the opinion system to be used to publicly criticize any user of the portal.
The challenge of architecting such a platform stems from two goals. First, leverage as much as possible of the rich functionality already provided by content management systems (CMS) and social networking platforms [15, 21]. For instance, there is no reason to re-implement user profiles. Second, the platform should be extensible, facilitate the development of software modules specific to preservation and archive extension, such as support for audio resources, and ensure their seamless communication with generic modules. Here, we discuss how we leveraged existing platforms and which of the current platform functionality we developed in-house. The platform is currently being developed in the context of NNAILA and the source code is publicly available [45] and will be updated as development proceeds.

NNAILA is based on Drupal [15], an open source and web-based content management platform. Drupal allows a community of users to publish, manage and organize a variety of content online and serves as the back-end for different types of websites ranging from small personal websites and blogs to large commercial web applications. Drupal’s modular and extensible architecture, and the numerous provided modules implementing the most common functionality expected of a web application, fit the goals of the Open Preservation and Extension Platform. Currently, the NNAILA Community Portal [44] leverages Drupal modules providing content management, indexing and search, user management (registration, authentication and profiling), web-based administration and support for content popularity analysis and user tracking. However, the unique preservation and archive extension functionality required by NNAILA has already necessitated the development of new Drupal modules, listed below, and we will continue to build on the system, as needed, for the present project:

- **Audio Content** The NNAILA Archive consists entirely of audio resources created by digitizing field recordings of Onondaga word elicitation, phrases, conversations and stories. Each audio resource is also annotated with metadata, such as the recording date, the speaker, etc. To publish these archive resources on the Community Portal, a new “Audio” content type, and supporting functionality, was developed. A web interface was implemented to upload and create audio resources. The interface analyzes new audio resources and automatically associates extracted features, such as play time, to metadata fields that are part of the Dublin Core Metadata Initiative (DCMI) terms [18].

- **User Contributions** Drupal provides a module supporting user discussions on published content, but contributions are limited to free text, an opinion mechanism is not incorporated, and discussions referring to multimedia resources, such as audio, are not supported. First, we enhanced the Drupal module to support user contributions on specific sections of an audio resource. An example is the first user contribution in Figure 2 referring to the initial 26 seconds of the audio resource, that can be played back in isolation within the space displaying the content of the contribution. Second, the module now supports rich user contributions, such as PDF, image and audio files. A third enhancement is the opinion mechanism incorporated and displayed within every user contribution.

- **Semantic Relationships** Archive resources have semantic relationships with each other. For example, relationship type *Same Session As* denotes that resources were recorded during the same session (Figure 2), but have been separated because of other considerations, such as the tape had to be changed, a single audio resource would be too long, etc. To support arbitrary relationships between archive resources, a module was developed as part of the Community Portal. The module enables the Archivist to define new relationship types and edit existing ones, and establish relationships between any pair of archive resources. For a given archive resource, the module also renders related resources and the connecting relationship types, as shown in the upper right corner of Figure 2.

- **Sophisticated Search and Filtering** The out of the box content search functionality provided by Drupal was not exploiting the rich metadata of NNAILA’s audio resources. Searching archive resources

---

9 Note that the need for platform extensibility does not align well with popular social networking platforms that follow the web hosted model [43, 28].

11
by a speaker’s name or filtering search results for a given field researcher should be possible. Hence, we enhanced Drupal’s module to support search on all metadata fields for a given set of keywords, filtering of the search results by a combination of metadata field values, and sorting of the filtered search results by several options, such as sort by creation date.

6 Needs Assessment
Parallel with the technical aspects of the project, work will be done on a needs assessment for NNAILA. Two broad areas will be the focus of this assessment: (i) the state of documentary resources on northeastern North American languages and (ii) possible uses for these materials for which there is the most demand.

With respect to resources, key questions to be considered are: (i) What documentary resources exist on the languages of Northeastern North America? (ii) Of these resources, which are already being housed in environments which offer reasonable assurances for their long-term preservation? (iii) What is the extent of digitized resources already available? and (iv) What high-value resources seem most in need of digitization and archiving (and, more generally, how should digitization activities be prioritized)?

With respect to uses of the materials, the key questions will be: (i) What sorts of functionalities would best facilitate the use of these materials by the Onondaga? (ii) To what extent are the needs of the Onondaga representative of those of other northeastern North American communities and, in cases where they are not, what other functionalities should be prioritized? (iii) What kinds of materials will be of most utility to speaker and research communities and to what extent should this impact digitization prioritization? and (iv) What general principles should NNAILA adopt to assist in the process through which resource sensitivities are assessed for the Onondaga materials and beyond?

Given that the resources housed in NNAILA are of interest both to the research community and to speaker communities, an obvious question is why the speaker community is being prioritized here over the research community. Our choice here is largely driven by the fact that language archives have generally found that members of speaker communities constitute a much larger share of their uses than researchers [5]. Therefore, focusing on this community is not only likely to benefit more people, in the context of the present project, it should also offer as an opportunity to work with a larger and more engaged community than if our initial focus was the research community.

Products resulting from this needs assessment are discussed in Section 7. In the context of the present project, a key result of this needs assessment will be to allow us to arrive at a better understanding regarding how the work described here can be effectively extended to a wider range of user communities.

7 Research Plan, Evaluation, and Project Deliverables
7.1 Research Plan
We break down the overall plan of research by half-year periods, dividing the work into computational (C) and linguistic (L) portions, assuming a start date of June 2010. Each investigator will be responsible with overseeing the portions of the work matching their expertise and will be assisted by a graduate student.

The primary responsibilities of the computer science graduate student will be to (i) design specifications of new functionality, communicate them to the linguistic side of the project, and adjust design based on feedback, (ii) propose and implement solutions to the open issues listed in Section 4 related to publishing archive resources, promoting user contributions and identifying quality community resources and reliable users, (iii) develop evaluation frameworks for the proposed solutions, possibly by monitoring the social interaction on the Community Portal, (iv) maintain the NNAILA development, testing and production sites, and (v) package and distribute the developed software online.

The responsibilities of the linguistics graduate student will be to (i) coordinate communication between the computational side of the project and the indigenous speaker communities to be engaged with over the
course of the work, (ii) collect information as needed for the needs assessment discussed in Section 6, (iii) monitoring of the contributions made to the developing social network and relaying relevant observations to the rest of the project team, (iv) hand annotation of a set of community contributions to facilitate development of the quality identification algorithms to be developed, (v) evaluation of ability of quality-determining algorithms to find high-quality contributions, and (vi) evaluation of the usability of the system for contribution promotion. Additionally, as time allows, they will also be asked to digitize materials to be deposited in the archive, with an emphasis on materials which will most directly further the specific aims of this project.

Timeline

- **Jun. 2010**
  - C: Development of a resource-oriented role-based access control mechanism, implementation of support for multiple collections of archive resources, and design of the web-based promotion interface for the Archivist.
  - L: Initiation of contacts with resource holders of northeastern North American Native Language materials as part of needs assessment.

- **Jan. 2011**
  - C: Implementation of the promotion interface and the archive correction mechanism for the community users to report archive errors and for the Archivist to correct them. Design of two software components capturing social interaction and promotion provenance information.
  - L: Initial compilation of known language resources, including preliminary prioritization. (Products of this work will be continuously refined over the course of the grant.) Evaluation of user interface for platform in preparation for community testing.

- **Jun. 2011**
  - C: Development of components capturing social interaction and promotion provenance information, and design of identification algorithms for quality contributions and reliable community users based on them.\(^{10}\) Implementation of interfaces for submitting user contributions to archive resources of type text and image.
  - L: Coordination between the project team and Onondaga community members, in collaboration with Dr. Abrams (see letter of commitment), in testing and evaluating the developing platform with a focus on newly implemented functionalities and the extent of community adoption.

- **Jan. 2012**
  - C: Development, deployment and testing of the first version of the identification algorithms and corresponding evaluation framework. Design and implementation of audio snipping.
  - L: Cooperation with computer science half of team to refine system in ways consistent with first evaluation.

- **Jun. 2012**
  - C: Design of alternative opinion mechanisms for the second version of the identification algorithms. Implementation of support for rich semantic relationships among archive resources.
  - L: Coordination between the project team and Onondaga community members in testing and evaluating the developing platform with a focus on newly implemented functionalities and the extent of community adoption of the refined system.

- **Jan. 2013**
  - C: Development and evaluation of the second version of the identification algorithms. Implementation of support for offline batch upload and for propagating semantic web graph extensions to the Archive in the long-term preservation storage.

\(^{10}\) We should make clear here that data relating to quality of contributions and reliability of users will be made directly available only to individuals involved in the archiving process and will not be “published” on the portal.
7.2 Evaluation

In addition to biweekly project meetings during the academic year and during the summer as needed, formal evaluation of project progress will take place along two tracks. One of these will assess the social side of the project in order to determine the extent to which the community of focus is able to make use of the system in ways which allow the activities of individual users to produce contributions of high quality. The other will assess the developing software regarding what aspects of the overall design plan have been adequately implemented and the extent to which those implementations appropriately address the special sensitivities and needs of a language archive.

The mechanisms for social evaluation will be: (i) determination of the extent to which those aspects of user interaction captured by the portal permit the detection of high-quality contributions, (ii) determination of the extent to which user contributions contain information judged of sufficient value to be promoted, (iii) informal observation of the use of the system in situations both where members of the research system are present with users and also via periodic post hoc evaluation of user activities, and (iv) the use of user surveys to solicit feedback on specific aspects of the portal. (In the model in Figure 4 discussed in more detail immediately below, these evaluation activities will take place during the “Social evaluation” phase.)

The scheme for software evaluation is laid out in Figure 4. It has become clear from experiences already gained from project development that it is essential members of both speaker communities (in this case, the Onondaga community) and the researcher community (in this case, primarily linguists) need to be brought in at the stage of overall functionality specification. While early input from such crucial stakeholders in a project like this is always valuable, the particular sensitivities surrounding the materials in an archive like NNAILA make having it especially important since even the most well-intentioned research team may fail to anticipate important sensitivities.

In addition to specification of system functionality, software development is broken down into three other components. The first is an initial implementation supervised primarily by the computer science project personnel. The second is an evaluation of this initial implementation from which input will be specifically solicited from the linguistics project personnel and researchers (including Onondaga researchers). The final component is evaluation of “live” services made available to members of the speaker community.

Finally, the customizability of the system as described in Section 5 will allow us to concretely test how changes to user interface may affect the extent to which users are able to interact with the system in ways which produce quality contributions. For example, perhaps the use of different schemes for user-provided
ratings of other users’ contributions will have a significant impact on the system’s ability to determine a contributions’ quality.\textsuperscript{11}

7.3 Project Deliverables
Project deliverables can be broken down into (i) models and methods, (ii) tools, and (iii) resources. Each of these is addressed in turn.

Models and to be developed by the project include those for:
- A community of interest centered around a resource
- Detection of quality of community contributions associated with a resource
- Promotion of community contributions to archival objects

Tools to be developed by the project include those for:
- A social networking system optimized for communities of interest based around resources
- Coordination between a social networking representation of an object and an archival object
- Migration of community contributions to a form which can be deposited within an archive

Resources to be developed by the project:
- An inventory of known documentary resources of northeastern North American languages
- A document suggesting prioritization of work on archiving materials on these languages based on diverse factors including need for preservation, community interest, and overall feasibility

8 Education Plan and Broader Impacts

Education plan: The modular architecture of the proposed platform will be exploited in teaching, tying together various courses from the standard computer science and linguistics curriculum. For instance, the current UB Web Data Engineering (CSE510) course uses a social networking platform for the course project on developing a portal. Students will be able to publish their own archive and enrich it with publicly-provided resources. The underlying algorithms on the other hand are of interest to the advanced Data Integration (CSE636) graduate course. The platform’s functionality will also allow it play a role in the Linguistics Department’s Field Methods course (LIN527), as a repository for resources collected by students in the class. In year three of the grant, project PIs will investigate the feasibility of teaching a joint course advertised to computer science students, linguists, and students interested in the digital humanities which will explore how well the platform can be adopted to other domains (e.g., digitized items from UB Libraries Special Collections).

Broader impacts: The proposed research will develop and distribute infrastructure enabling the controlled sharing of data on the Web and the formation of communities of interest around high-quality resources. The software, methods, and models to be developed by this project will be of potential use to any project seeking to facilitate resource-based communities, and, because of its thematic focus on language materials, will be of particular use for digital archives of language resources. In the context of Onondaga, the project will be able to play a role in giving Onondaga community members access to their heritage in ways that allow them to shape the record of their heritage themselves. This will result in new knowledge about Onondaga and also may have a positive impact on the use of the Onondaga language in community, thereby helping to ensure it is passed on to future generations.

\textsuperscript{11} Where the opportunity arises to examine such questions, the project team will consult with outside researchers and community members following the model in Figure 4 to see how this can be done in ways which properly take into account any community sensitivities, since it is crucial that a project like this one avoid adopting an overly “experimental” approach which might disincline people from participating in tool development.
References


Biographical Sketch

Michalis Petropoulos
Assistant Professor
Computer Science & Engineering Department
University at Buffalo, SUNY
201 Bell Hall, Box 602000
Buffalo, NY 14260-2000

mpetropo@buffalo.edu
http://www.cse.buffalo.edu/~mpetropo
Phone: (716) 645-4733
Fax: (716) 645-3464

PROFESSIONAL PREPARATION

Technical University of Crete, Greece
Computer Engineering
Diploma 1998

University of California, San Diego
Computer Science
MS 2000

University of California, San Diego
Computer Science
PhD 2005

APPOINTMENTS

University at Buffalo, SUNY, Buffalo, NY
Jan 2006 – present
Assistant Professor, Computer Science & Engineering Department

App2you, Inc., La Jolla, CA
Jan – August 2007
Consultant

University of California, San Diego, La Jolla, CA
Sep 1999 – Dec 2005
Research Assistant, Computer Science & Engineering Department

IBM Almaden Research Center, San Jose, CA
Jun - Sept 2004
Research Intern, Advanced Database Services Group

Microsoft Research, Redmond, WA
Jul – Nov 2003
Research Intern, Database Group

Bell Laboratories, Murray Hill, NJ
Jun - Sep 2002
Research Intern, Network Data and Services Research Department

Enosys Software Inc., San Diego, CA
Jan 2000 – Sep 2001
Senior Software Engineer

PUBLICATIONS MOST RELATED TO PROPOSAL

1. BioNav: Effective Navigation on Query Results of Biomedical Databases
   Abhijith Kashyap, Vagelis Hristidis, Michalis Petropoulos, Sotiria Tavoulari
   International Conference on Data Engineering (ICDE), 2009

2. Exporting and Interactively Querying Web Service-Accessed Sources: The CLIDE System
   Michalis Petropoulos, Alin Deutsch, Yannis Papakonstantinou, Yannis Katsis
   ACM Transactions on Database Systems (TODS), 32(4), 2007

3. Interactive Query Formulation over Web Service-Accessed Sources
   Michalis Petropoulos, Alin Deutsch, Yannis Papakonstantinou
   ACM International Conference on Management of Data (SIGMOD), 2006

4. Graphical Query Interfaces for Semistructured Data: The QURSED System
   Michalis Petropoulos, Yannis Papakonstantinou, Vasilis Vassalos
   ACM Transactions on Internet Technology (TOIT), 5(2), 2005

5. QURSED: Querying and Reporting SEMistructured Data
   Yannis Papakonstantinou, Michalis Petropoulos, Vasilis Vassalos
   ACM International Conference on Management of Data (SIGMOD), 2002
OTHER PUBLICATIONS

1. **Do-It-Yourself Custom Forms-Driven Workflow Applications**  
   Keith Kowalczykowski, Kian Win Ong, Alin Deutsch, Yannis Papakonstantinou, Kevin Zhao, Michalis Petropoulos  
   Conference on Innovative Data Systems Research (CIDR), 2009

2. **Industrial-Strength Schema Matching**  
   Philip A. Bernstein, Sergey Melnik, Michalis Petropoulos, Christoph Quix  
   ACM SIGMOD Record, 33(4), 2004

3. **Query Set Specification Language (QSSL)**  
   Michalis Petropoulos, Alin Deutsch, Yannis Papakonstantinou  
   International Workshop on the Web and Databases (WebDB), 2003

4. **Semantic Caching of XML Databases**  
   Vagelis Hristidis, Michalis Petropoulos  
   International Workshop on the Web and Databases (WebDB), 2002

5. **Building XML Query Forms and Reports with XQForms**  
   Michalis Petropoulos, Yannis Papakonstantinou, Vasilis Vassalos  

SYNERGISTIC ACTIVITIES

**Co-Chair:**  
WebDB 2009 – International Workshop on the Web and Databases

**Tutorials Chair:**  
ICWE 2007 – International Conference on Web Engineering

**PC Chair:**  
NTII 2010 - International Workshop on New Trends in Information Integration  
WWW 2010 - International World Wide Web Conference  
EDBT 2010 - International Conference on Extending Database Technology

RECENT COLLABORATORS

- Alin Deutsch – University of California, San Diego (CS Dept)  
- Jeff Good – University at Buffalo, SUNY (Linguistics Dept)  
- Vagelis Hristidis – Florida International University (CS Dept)  
- Hung Quang Ngo – University at Buffalo, SUNY (CS Dept)  
- Sotiria Tavoulari – Yale University (Pharmacology Dept)  
- Shambhu Upadhyaya – University at Buffalo, SUNY (CS Dept)

GRADUATE AND POSTDOCTORAL ADVISORS

- Graduate Advisor: Yannis Papakonstantinou (UC San Diego)

THESIS ADVISOR AND POSTGRADUATE-SCHOLAR SPONSOR

**PhD Students (2 current):**  
- Demian Lessa (University at Buffalo, SUNY)  
- Abhijith Kashyap (University at Buffalo, SUNY)
Biographical sketch

Jeff Good

Professional preparation

University of Chicago, Chicago, Illinois
• B.A., Linguistics, 1998
• M.A., Linguistics, 1998

University of California, Berkeley, Berkeley, California
• M.A., Linguistics, 2000
• Ph.D., Linguistics, 2003

Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
• Postdoctoral fellow, Department of Linguistics, 2004–2005.

Appointments

2006–Pres. University at Buffalo, Department of Linguistics
    Assistant Professor
2005–2006 Max Planck Institute for Evolutionary Anthropology, Department of Linguistics
    Senior Research Fellow
2003–2004 University of Pittsburgh, Department of Linguistics
    Visiting Assistant Professor

Related publications

(Pre-publication copies are available at http://buffalo.edu/~jcgood/publications.html for works not yet published and some published works.)


Other publications
(Pre-publication copies are available at http://buffalo.edu/~jcgood/publications.html for works not yet published and some published works.)


Synergistic activities
2002–Pres. Chair, Outreach Working Group of Open Language Archives Community
2006–Pres. Technical director, Rosetta Project, San Francisco, California
2006–Pres. Chair, Linguistic Society of America (LSA) Technology Advisory Committee
2009 Invited to teach *Data management for field linguists* at the LSA Summer Institute
2009–2010 Interim director, Digital Humanities Initiative at Buffalo, University at Buffalo

Collaborators and co-editors
Anthony Aristar (Eastern Michigan University), Helen Aristar-Dry (Eastern Michigan University), Emily M. Bender (University of Washington), Hans-Jörg Bibiko (MPI EVA), Gene Buckley (University of Pennsylvania), Laura Buszard-Welcher (Long Now Project), Denis Creissels (University of Lyon), Michael Cysouw (MPI EVA), Eric Delson (Lehman College, CUNY), Lise Dobrin (University of Virginia), Carol Ember (Human Relations Area Files), Scott Farrar (University of Arizona), Michael D. Fischer (University of Kent), Tom Gündeman (MPI EVA), Larry Hyman (UC Berkeley), Heidi Johnson (UT Austin), Karin Michelson (University at Buffalo), John McWhorter (Manhattan Institute), Sebastian Nordhoff (University of Amsterdam), Shakhti Poornima (University at Buffalo), Carolyn O’Meara (University at Buffalo), Michalis Petropoulos (University at Buffalo), Stuart Robinson (Powerset), Dean Snow (Pennsylvania State University)

Graduate and postdoctoral advisors
Graduate Advisor: Larry Hyman (UC Berkeley)
Postdoctoral Advisor: Bernard Comrie (MPI EVA)

Thesis advisor and postgraduate-scholar sponsor (4 graduate students)
Joseph T. Farquharson (University of the West Indies), Cristin Kalinowski (University at Buffalo), Jesse Lovegren (University at Buffalo), Nyurguyana Petrova (University at Buffalo), Rebecca Voll (University of Leiden)
<table>
<thead>
<tr>
<th>ORGANIZATION</th>
<th>SUNY at Buffalo</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR</td>
<td>Michail Petropoulos</td>
</tr>
<tr>
<td>PROPOSAL BUDGET</td>
<td>FOR NSF USE ONLY</td>
</tr>
<tr>
<td>SUMMARIZED PROPOSAL BUDGET</td>
<td>YEAR 1</td>
</tr>
</tbody>
</table>

**A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates**

(List each separately with title, A.7. show number in brackets)

<table>
<thead>
<tr>
<th>Person</th>
<th>CAL</th>
<th>ACAD</th>
<th>SUMR</th>
<th>Funds Requested By</th>
<th>Funds Granted by NSF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Michail Petropoulos - PI</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>$10,954</td>
<td>$</td>
</tr>
<tr>
<td>2. Jeffrey Good - Co-I</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>$7,559</td>
<td>$</td>
</tr>
</tbody>
</table>

**B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS)**

<table>
<thead>
<tr>
<th>Category</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. POST DOCTORAL SCHOLARS</td>
<td>(        )</td>
</tr>
<tr>
<td>2. OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.)</td>
<td>(        )</td>
</tr>
<tr>
<td>3. GRADUATE STUDENTS</td>
<td>(        )</td>
</tr>
<tr>
<td>4. UNDERGRADUATE STUDENTS</td>
<td>(        )</td>
</tr>
<tr>
<td>5. SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY)</td>
<td>(        )</td>
</tr>
<tr>
<td>6. OTHER</td>
<td>(        )</td>
</tr>
</tbody>
</table>

**C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)**

**D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING $5,000.)**

See budget justification

**E. TRAVEL**

<table>
<thead>
<tr>
<th>Category</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. DOMESTIC (INCL. CANADA, MEXICO AND U.S. POSSESSIONS)</td>
<td>(        )</td>
</tr>
<tr>
<td>2. FOREIGN</td>
<td>(        )</td>
</tr>
</tbody>
</table>

**F. PARTICIPANT SUPPORT COSTS**

<table>
<thead>
<tr>
<th>Category</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>STIPENDS</td>
<td>$0</td>
</tr>
<tr>
<td>TRAVEL</td>
<td>$0</td>
</tr>
<tr>
<td>SUBSISTENCE</td>
<td>$0</td>
</tr>
<tr>
<td>OTHER</td>
<td>$0</td>
</tr>
</tbody>
</table>

**G. OTHER DIRECT COSTS**

<table>
<thead>
<tr>
<th>Category</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>MATERIALS AND SUPPLIES</td>
<td>$0</td>
</tr>
<tr>
<td>PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION</td>
<td>$0</td>
</tr>
<tr>
<td>CONSULTANT SERVICES</td>
<td>$1,500</td>
</tr>
<tr>
<td>COMPUTER SERVICES</td>
<td>$4,500</td>
</tr>
<tr>
<td>SUBAWARDS</td>
<td>$0</td>
</tr>
<tr>
<td>OTHER</td>
<td>$19,584</td>
</tr>
</tbody>
</table>

**H. TOTAL DIRECT COSTS (A THROUGH G)**

**I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE)**

**MTDC (Rate: 58.5000, Base: 82196)**

Total indirect costs (F&A) $48,085

**J. TOTAL DIRECT AND INDIRECT COSTS (H + I)**

$149,864

**K. RESIDUAL FUNDS**

$0

**L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K)**

$149,864

**M. COST SHARING PROPOSED LEVEL $**

**Not Shown**

**AGREED LEVEL IF DIFFERENT $**

**PI/PD NAME**

Michail Petropoulos

**FOR NSF USE ONLY**

**ORG. REP. NAME**

Jeffrey Schieder

*ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET*
## SUMMARY PROPOSAL BUDGET

**Organization**: SUNY at Buffalo

### Principal Investigator / Project Director
- **Michail Petropoulos**

### Proposal Budget

<table>
<thead>
<tr>
<th>A. Senior Personnel: PI/PD, Co-PI’s, Faculty and Other Senior Associates</th>
<th>NSF Funded</th>
<th>ACAD</th>
<th>SUMR</th>
<th>Funds Requested by proposer</th>
<th>Funds granted by NSF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Michail Petropoulos - PI</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>$11,502</td>
<td>$</td>
</tr>
<tr>
<td>2. Jeffrey Good - Co-I</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. (0) Others (List individually on Budget Justification Page)</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7. (2) Total Senior Personnel (1-6)</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>$11,502</td>
<td>0</td>
</tr>
</tbody>
</table>

### B. Other Personnel (Show numbers in brackets)

<table>
<thead>
<tr>
<th>B. Other Personnel</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. (0) Post Doctoral Scholars</td>
</tr>
<tr>
<td>2. (0) Other Professionals (Technician, Programmer, Etc.)</td>
</tr>
<tr>
<td>3. (2) Graduate Students</td>
</tr>
<tr>
<td>4. (0) Undergraduate Students</td>
</tr>
<tr>
<td>5. (0) Secretarial - Clerical (If charged directly)</td>
</tr>
<tr>
<td>6. (0) Other</td>
</tr>
</tbody>
</table>

**Total Salaries and Wages (A + B)**: 61,902

### C. Fringe Benefits (If charged as Direct Costs)

**Total Salaries, Wages and Fringe Benefits (A + B + C)**: 71,302

### D. Equipment (List item and dollar amount for each item exceeding $5,000.)

**Total Equipment**: 0

### E. Travel

- **Domestic (Incl. Canada, Mexico and U.S. Possessions)**: 2,400
- **Foreign**: 0

### F. Participant Support Costs

<table>
<thead>
<tr>
<th>F. Participant Support Costs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Stipends</td>
</tr>
<tr>
<td>2. Travel</td>
</tr>
<tr>
<td>3. Subsistence</td>
</tr>
<tr>
<td>4. Other</td>
</tr>
</tbody>
</table>

**Total Number of Participants**: 0

**Total Participant Costs**: 0

### G. Other Direct Costs

- **Materials and Supplies**: 0
- **Publication Costs/Documentation/Dissemination**: 0
- **Consultant Services**: 4,500
- **Computer Services**: 4,500
- **Subawards**: 0
- **Other**: 19,584

**Total Other Direct Costs**: 28,084

### H. Total Direct Costs (A through G)

**Total Direct Costs**: 101,786

### I. Indirect Costs (F&A) (Specify Rate and Base)

**MTDC (Rate: 58.5000, Base: 82203)**

**Total Indirect Costs (F&A)**: 48,089

### J. Total Direct and Indirect Costs (H + I)

**Total Direct and Indirect Costs**: 149,875

### K. Residual Funds

**Residual Funds**: 0

### L. Amount of This Request (J) or (J Minus K)

**Amount**: 149,875

### M. Cost Sharing Proposed Level

**Cost Sharing Proposed Level**: Not Shown

**Agreed Level if Different**: $0

### PI/PD Name

- **Michail Petropoulos**

### Org. Rep. Name

- **Jeffrey Schieder**

---

2 *ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET*
## Proposal Budget

<table>
<thead>
<tr>
<th>Category</th>
<th>Amount</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A. Senior Personnel</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Michail Petropoulos - PI</td>
<td>8,334</td>
<td></td>
</tr>
<tr>
<td>Jeffrey Good - Co-I</td>
<td>8,334</td>
<td></td>
</tr>
<tr>
<td><strong>B. Other Personnel</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post Doctoral Scholars</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Other Professionals</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Graduate Students</td>
<td>52,920</td>
<td></td>
</tr>
<tr>
<td>Undergraduate Students</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Secretarial/Clerical (if charged directly)</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><strong>C. Fringe Benefits</strong></td>
<td>9,536</td>
<td></td>
</tr>
<tr>
<td><strong>D. Equipment</strong></td>
<td>70,790</td>
<td></td>
</tr>
<tr>
<td><strong>E. Travel</strong></td>
<td>2,400</td>
<td></td>
</tr>
<tr>
<td>Domestic (incl. Canada, Mexico and U.S. possessions)</td>
<td>2,400</td>
<td></td>
</tr>
<tr>
<td>Foreign</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><strong>F. Participant Support Costs</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stipends</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Travel</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Subsistence</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><strong>G. Other Direct Costs</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Materials and Supplies</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Publication Costs/Documentation/Dissemination</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Consultant Services</td>
<td>4,500</td>
<td></td>
</tr>
<tr>
<td>Computer Services</td>
<td>4,500</td>
<td></td>
</tr>
<tr>
<td>Subawards</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>19,584</td>
<td></td>
</tr>
<tr>
<td><strong>H. Total Direct Costs</strong></td>
<td>101,774</td>
<td></td>
</tr>
<tr>
<td><strong>I. Indirect Costs (F&amp;A) (Specify Rate and Base)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MTDC (Rate: 58.5000, Base: 82190)</td>
<td>149,855</td>
<td></td>
</tr>
<tr>
<td><strong>J. Total and Indirect Costs (H + I)</strong></td>
<td>149,855</td>
<td></td>
</tr>
<tr>
<td><strong>K. Residual Funds</strong></td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><strong>L. Amount of This Request (J) or (J minus K)</strong></td>
<td>149,855</td>
<td></td>
</tr>
<tr>
<td><strong>M. Cost Sharing</strong></td>
<td>Not Shown</td>
<td>Agreed level if different</td>
</tr>
</tbody>
</table>

### Summary

**Organizations**
- SUNY at Buffalo
- SUNY at Buffalo
- SUNY at Buffalo

**Principal Investigator / Project Director**
- Michail Petropoulos

**Award Number**
- 000000

**Proposal No.**
- Proposed

**Duration (months)**
- Proposed

**Funds Requested by Proposer**
- 0

**Funds Granted by NSF**
- 0

**Date Checked**
- 3

**Date Of Rate Sheet**
- 3

**Initials - ORG**
- ORG

---

3. *Electronic signatures required for revised budget*
### SUMMARY

#### PROPOSAL BUDGET

**FOR NSF USE ONLY**

<table>
<thead>
<tr>
<th>ORGANIZATION</th>
<th>SUNY at Buffalo</th>
</tr>
</thead>
</table>

| PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR | Michail Petropoulos |

#### A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates

(List each separately with title, A.7. show number in brackets)

<table>
<thead>
<tr>
<th>Name</th>
<th>CAL</th>
<th>ACAD</th>
<th>SUMR</th>
<th>NSF Funded</th>
<th>Person-months</th>
<th>Funds Requested by proposer</th>
<th>Funds granted by NSF (if different)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Michail Petropoulos - PI</td>
<td>0.00</td>
<td>0.00</td>
<td>2.00</td>
<td>$22,456</td>
<td>$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Jeffrey Good - Co-I</td>
<td>0.00</td>
<td>0.00</td>
<td>2.00</td>
<td>$15,893</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS)

1. (0) POST DOCTORAL SCHOLARS
2. (0) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.)
3. (6) GRADUATE STUDENTS
4. (0) UNDERGRADUATE STUDENTS
5. (0) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY)
6. (0) OTHER

**TOTAL SALARIES AND WAGES (A + B)**: $189,669

#### C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)

**TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C)**: $218,287

#### D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING $5,000.)


#### E. TRAVEL

1. DOMESTIC (INCL. CANADA, MEXICO AND U.S. POSSESSIONS)
2. FOREIGN

#### F. PARTICIPANT SUPPORT COSTS

1. STIPENDS
2. TRAVEL
3. SUBSISTENCE
4. OTHER

**TOTAL NUMBER OF PARTICIPANTS**

**TOTAL PARTICIPANT COSTS**

#### G. OTHER DIRECT COSTS

1. MATERIALS AND SUPPLIES
2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION
3. CONSULTANT SERVICES
4. COMPUTER SERVICES
5. SUBAWARDS
6. OTHER

**TOTAL OTHER DIRECT COSTS**: $82,252

#### H. TOTAL DIRECT COSTS (A THROUGH G)

**TOTAL DIRECT COSTS**: $305,339

#### I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE)

**TOTAL INDIRECT COSTS (F&A)**: $144,255

#### J. TOTAL DIRECT AND INDIRECT COSTS (H + I)

**TOTAL DIRECT AND INDIRECT COSTS**: $449,594

#### K. RESIDUAL FUNDS

**RESIDUAL FUNDS**

#### L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K)

**AMOUNT OF THIS REQUEST**: $449,594

#### M. COST SHARING PROPOSED LEVEL

**COST SHARING PROPOSED LEVEL**: Not Shown

**AGREED LEVEL IF DIFFERENT**: $449,594

---

**PI/PD NAME**

Michail Petropoulos

**INDIRECT COST RATE VERIFICATION**

---

**FOR NSF USE ONLY**

---

**ORG. REP. NAME**

Jeffrey Schieder

---

C *ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET*
Budget Justification

A. Senior personnel

- Michalis Petropoulos (PI): One summer month in year one and one summer month year two (and other effort as needed) for supervision of overall project with Co-PI Good and for primary supervision of its computer science component.

- Jeff Good (Co-PI): One summer month in year one and one summer month in year three (and other effort as needed) for supervision of overall project with PI Petropoulos and for primary supervision of its linguistics component.

B. Other personnel

- Two graduate student researchers, one from Department of Computer Science and Engineering and one from Department of Linguistics: $24000 two-semester salary for each in year one, $25200 two-semester salary for each in year two, and $26460 for each in year three.

C. Fringe Benefits

- Fringe benefits are budgeted at the federally negotiated rates.

E. Travel

- Six domestic trips to present results of research: $800 each.

G. Other Direct Costs

G3. Consultant services

- Consultant fees to Dr. Percy Abrams for serving as liaison between research team and Onondaga community at Onondaga Nation: $1500 in year one of project (30 hrs. @ $50/hr.), $2500 (50 hrs. @ $50/hr.) in year two, and $3000 (60 hrs. @ $50/hr.) in year three.

- Consultant fees to linguists with materials deposited at NNAILA for testing the tools to be developed: $750 (15 hrs. @ $50/hr.) in year two and $750 (15 hrs. @ $50/hr.) in year three.

- Consultant fees to Onondagas (especially language teachers) for testing the tools to be developed: $750 (15 hrs. @ $50/hr.) in year two and $750 (15 hrs. @ $50/hr.) year three.

G4. Computer services

- Negotiated rate with Department of Computer Science and Engineering for computer services: $4500 for each year of grant.

G6. Other

- Three years of tuition for two University at Buffalo graduate student assistants: $19584 for each year of grant.

I. Indirect Costs

- Indirect costs are budgeted at the federally negotiated rate of 58.5% modified total direct costs.
### Current and Pending Support

(See GPG Section II.C.2.h for guidance on information to include on this form.)

The following information should be provided for each investigator and other senior personnel. Failure to provide this information may delay consideration of this proposal.

<table>
<thead>
<tr>
<th>Investigator: Michail Petropoulos</th>
<th>Other agencies (including NSF) to which this proposal has been/will be submitted.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support: ☐ Current ☑ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support</td>
<td></td>
</tr>
<tr>
<td>Project/Proposal Title: CAREER: Integrating and Interactively Querying Data Services</td>
<td></td>
</tr>
<tr>
<td>Source of Support: NSF</td>
<td></td>
</tr>
<tr>
<td>Total Award Amount: $581,375 Total Award Period Covered: 02/01/10 - 01/31/15</td>
<td></td>
</tr>
<tr>
<td>Location of Project: University At Buffalo, SUNY</td>
<td></td>
</tr>
<tr>
<td>Person-Months Per Year Committed to the Project. Cal: 0.00 Acad: 0.00 Sumr: 2.00</td>
<td></td>
</tr>
</tbody>
</table>

| Support: ☐ Current ☑ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support |
| Project/Proposal Title: MRI-R2: Acquisition of a Data Intensive Supercomputer for Innovative and Transformative Research in Science and Engineering |
| Source of Support: NSF |
| Total Award Amount: $4,600,351 Total Award Period Covered: 12/31/09 - 12/30/12 |
| Location of Project: University At Buffalo, SUNY |
| Person-Months Per Year Committed to the Project. Cal: 0.00 Acad: 0.00 Sumr: 0.00 |

| Support: ☐ Current ☑ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support |
| Project/Proposal Title: DEL/SoCS: Extending Preservation Archives through Social Networking |
| Source of Support: NSF |
| Total Award Amount: $449,594 Total Award Period Covered: 06/01/10 - 05/31/13 |
| Location of Project: University at Buffalo |
| Person-Months Per Year Committed to the Project. Cal: 0.00 Acad: 0.00 Sumr: 1.00 |

| Support: ☐ Current ☑ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support |
| Project/Proposal Title: |
| Source of Support: |
| Total Award Amount: $ |
| Total Award Period Covered: |
| Location of Project: |
| Person-Months Per Year Committed to the Project. Cal: Acad: Sumr: |

*If this project has previously been funded by another agency, please list and furnish information for immediately preceding funding period.*
The following information should be provided for each investigator and other senior personnel. Failure to provide this information may delay consideration of this proposal.

### Investigator: Jeffrey Good

#### Support: [ ] Current   [ ] Pending   [ ] Submission Planned in Near Future   [ ] *Transfer of Support

<table>
<thead>
<tr>
<th>Project/Proposal Title</th>
<th>Source of Support</th>
<th>Total Award Amount: $</th>
<th>Total Award Period Covered:</th>
<th>Location of Project</th>
<th>Person-Months Per Year Committed to the Project: Cal:</th>
<th>Acad:</th>
<th>Sumr:</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTEROP: Lexicon Enhancement via the GOLD Ontology (LEGO)</td>
<td>NSF</td>
<td>636,443</td>
<td>09/01/08 - 08/31/11</td>
<td>Eastern Michigan University and University at Buffalo</td>
<td>0.00</td>
<td>0.50</td>
<td>0.00</td>
</tr>
</tbody>
</table>

#### Support: [ ] Current   [ ] Pending   [ ] Submission Planned in Near Future   [ ] *Transfer of Support

<table>
<thead>
<tr>
<th>Project/Proposal Title</th>
<th>Source of Support</th>
<th>Total Award Amount: $</th>
<th>Total Award Period Covered:</th>
<th>Location of Project</th>
<th>Person-Months Per Year Committed to the Project: Cal:</th>
<th>Acad:</th>
<th>Sumr:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toward an Integrated Plan for Digital Preservation and Access to Primary Anthropological Data (AnthroDataDPA): A Four-Field Workshop</td>
<td>NSF and Wenner-Gren</td>
<td>72,775</td>
<td>04/01/09 - 03/31/10</td>
<td>Human Relations Area Files at Yale University</td>
<td>0.00</td>
<td>0.50</td>
<td>0.00</td>
</tr>
</tbody>
</table>

#### Support: [ ] Current   [ ] Pending   [ ] Submission Planned in Near Future   [ ] *Transfer of Support

<table>
<thead>
<tr>
<th>Project/Proposal Title</th>
<th>Source of Support</th>
<th>Total Award Amount: $</th>
<th>Total Award Period Covered:</th>
<th>Location of Project</th>
<th>Person-Months Per Year Committed to the Project: Cal:</th>
<th>Acad:</th>
<th>Sumr:</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTEROP: Digital ReUse and Interoperability of Data (DRUID) for Four-Field Anthropological Research</td>
<td>NSF and Wenner-Gren</td>
<td>746,395</td>
<td>01/01/10 - 12/31/12</td>
<td>Human Relations Area Files at Yale University</td>
<td>0.00</td>
<td>0.00</td>
<td>0.75</td>
</tr>
</tbody>
</table>

#### Support: [ ] Current   [ ] Pending   [ ] Submission Planned in Near Future   [ ] *Transfer of Support

<table>
<thead>
<tr>
<th>Project/Proposal Title</th>
<th>Source of Support</th>
<th>Total Award Amount: $</th>
<th>Total Award Period Covered:</th>
<th>Location of Project</th>
<th>Person-Months Per Year Committed to the Project: Cal:</th>
<th>Acad:</th>
<th>Sumr:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Towards an Areal grammar of Lower Fungom (Abar [mij], Fang [fak], Koshin [kid], Kung [kfl], Mbu? [muc], Mundabli [boe], and Naki [mff])</td>
<td>NSF</td>
<td>319,725</td>
<td>09/01/09 - 08/31/12</td>
<td>University at Buffalo</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

#### Support: [ ] Current   [ ] Pending   [ ] Submission Planned in Near Future   [ ] *Transfer of Support

<table>
<thead>
<tr>
<th>Project/Proposal Title</th>
<th>Source of Support</th>
<th>Total Award Amount: $</th>
<th>Total Award Period Covered:</th>
<th>Location of Project</th>
<th>Person-Months Per Year Committed to the Project: Cal:</th>
<th>Acad:</th>
<th>Sumr:</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEL/SoCS: Extending Preservation Archives through Social Networking</td>
<td>NSF</td>
<td>449,594</td>
<td>06/01/10 - 05/31/13</td>
<td>University at Buffalo</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

*If this project has previously been funded by another agency, please list and furnish information for immediately preceding funding period.*
FACILITIES, EQUIPMENT & OTHER RESOURCES

FACILITIES: Identify the facilities to be used at each performance site listed and, as appropriate, indicate their capacities, pertinent capabilities, relative proximity, and extent of availability to the project. Use "Other" to describe the facilities at any other performance sites listed and at sites for field studies. USE additional pages as necessary.

Laboratory:

Clinical:

Animal:

Computer: The Department of Linguistics has a computer dedicated to audio digitization. The Department of Computer Science and Engineering has two servers available for project development and testing. The Libraries at the University at Buffalo will provide space on a public server for the

Office: Project Investigators will be able to use their regular university office space. Graduate student assistants will be given (possibly shared) office space by their respective departments.

Other:

MAJOR EQUIPMENT: List the most important items available for this project and, as appropriate identifying the location and pertinent capabilities of each.

The Department of Linguistics has the following equipment to facilitate digitization: An Analog-to-Digital converter, a reel-to-reel cassette player, a compact cassette player, and a "power cleaner" surge protector.

OTHER RESOURCES: Provide any information describing the other resources available for the project. Identify support services such as consultant, secretarial, machine shop, and electronics shop, and the extent to which they will be available for the project. Include an explanation of any consortium/contractual arrangements with other organizations.
COMPUTER FACILITIES (continued):

"live/production" version of the NNAILA portal.
September 12, 2009

Dear Dr. Good and Dr. Michalis,

As Director of the Onondaga Language Program, I write this letter to express my willingness to serve as a consultant on the project "Extending Preservation Archives through Social Networking," if it receives funding from the National Science Foundation. In particular, I will be able to serve as a liaison between this project and the Onondaga Language Program and offer input regarding the functionality and user interfaces of the proposed tools to ensure that they will be able to serve the needs of language instruction at Onondaga Nation.

Sincerely,
Percy Abrams, Ph.D.
Director of Onondaga Language Program

[Signature]

Ongwadeyenstha' Ononda'gekha'