Plans for a Digital Data Infrastructure for Bridging the Word Gap

Intervention Researchers

Charles R. Greenwood, Anne Larson, Judith J. Carta, Dale Walker, & Alana Schnitz

Juniper Gardens Children's Project

University of Kansas

December 4, 2016

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

INTRODUCTION

Fueling the scientific achievements of the last twenty years has been geometric growth in the capacity to collect increasingly more data at less cost per unit of information (Bell, Hey, & Szalay, 2009). This advance is a product of the digital revolution with the advent of the web; personal computing devices; digital sensors of many types; other applications capable of real time data collection; computer networks; and supercomputing (Thanos, 2011). The result has been stunning developments that have improved the "resolution" of our perception of phenomena and/or provided an entirely new vision -- the ability to see and understand phenomena previously unknown. For example, the Hubble telescope in space outside the earth's atmosphere has dramatically increased the resolution in what can be seen by scientists in the far galaxy, while the advent of digital mapping technology with heat mapping has created an entirely new lens for understanding the geographical diversity of weather systems, population density, etc., providing an entirely new perception of these phenomena.

With all of these developments have come an increasing need for new processes and systems to store, document, standardize, search, and use "big data" in order to address exciting new research questions. The networked Digital Data Infrastructure (DDI) has emerged as the basic tool through which both individual scientists and communities of collaborating scientists today are able to conduct their research in ways that both use and build on rapidly advancing knowledge and data.

> "*Research Data Infrastructures can be defined as managed networked environments for digital research data consisting of services and tools that support: (i) the whole research cycle, (ii) the movement of scientific data across scientific disciplines, (iii) the creation of open linked data spaces by connecting data sets from diverse disciplines, (iv) the management of scientific workflows, (v) the interoperation between scientific data and literature, and (vi) an Integrated Science Policy Framework.*" (Thanos, 2011, pg. 3).

> "*Research data infrastructures are not systems in the traditional sense of the term; they are networks that enable locally controlled and maintained digital data and library systems to interoperate more or less seamlessly*" (Thanos, 2011, pg. 3-4).

Using a DDI, a scientist may access datasets that can be used to address new research questions. Scientists collecting data based on common standards and documentation can contribute new data to the DDI for use by others. One can also post current research reports to the DDI, as well as a coding surveys describing articles' technical features for purposes of research syntheses by others. A graduate student can find, download, and use a data collection template developed by another researcher in the field for collecting dissertation data or conducting other studies that fulfill degree and course requirements.

The scope of DDI capabilities currently range from those serving individual communities of scientists within a particular scientific discipline f to a common topic or problem with global implications (e.g., climate change, etc.). DDI capabilities have progressed at differential speeds depending on disciplines and technological advances in discipline data and science. DDI developments have influenced nearly all scientific disciplines, and the looming promise of new discovery through using DDIs is on us.

The community of naturalistic language researchers interested in the 30 Million Word Gap is no exception. A fundamental example has been development of the LENA, a digital sensor that produces up to 16 hours of audio environment data experienced by an individual infant in a single day. The sensor (the LENA DLP) is a small audio recorder worn by an infant/toddler positioned in a custom designed chest pocket in the child's clothing (Greenwood et al., Submitted). The data stored on the device is uploaded to a PC or the Cloud, where speech

recognition software is used to produce automated reports on the quality of the child's audio environment and three standard language indicators: Adult Word Count, Conversational Turns, and Child Word Count (Richards, Gilkerson, Paul, & Xu, 2008; Xu, Yapanel, & Gray, 2009). The language indicators follow the coding taxonomy developed originally by Hart and Risley (1995).

The LENA and other nonintrusive, digital audio recording techniques used by the community of natural language researchers are developing DDIs containing naturalistic, day-long, digital audio recordings of young children and their audio environments (home, child care, preschool) that can used to support research. Speech recognition, linguistics, and developmental psychology scientists use the HOMEBANK DDI for sharing of day-long (http://homebank.talkbank.org/) audio records for research analysis as well as develop new methods of analyzing this data to discovery basic language processes.

The DATABRARY Project (https://databrary.org/about/mission.html), another DDI is devoted to archiving video information that can be contributed to and/or used in developmental research and related language and psychology fields. While these two DDIs are designed to serve the specific needs of developmental, linguistic, and speech scientists interested in descriptive, exploratory research that discovers language patterns and mechanisms in big data, these DDIs fall short of serving the unique needs of the community of language intervention researchers.

The language research community is seeking to prevent the adverse effects of the low SES related Word Gap as well as clinical applications to language intervention for subpopulations of children with language delays/disabilities. In contrast to other naturalistic language researchers, language intervention researchers' needs are uniquely focused on designing, conducting, and analyzing the results of experiments seeking to make improvements. While interests of interventionists overlap with those of speech and developmental scientists, a DDI serving intervention researchers will need to contain largely different resources.

The Language Intervention Research Community

The community of language intervention researchers and members of the Bridging the Word Gap Research Network (BWGRN) are presently focused on testing and validating alterable solutions to the problem of limited oral vocabulary learning experienced disproportionately by low SES infants, toddlers, and preschoolers (Carta, Greenwood, Walker, & Larson, 2016, February; Radesky, Carta, & Bair-Merritt, 2016). Low vocabulary by age 3 years is associated with not being ready for school and failure to learn to read in the elementary grades (Fernald, Marchman, & Weisleder, 2013). This disparity in child vocabulary size between low-versus moderate/high SES groups is traced to low exposure to talk in children's home and child care environments (Warren, 2015). This gap in exposure to words translates to a deficit in vocabulary growth that increases over time (Fernald, Marchman, & Weisleder, 2013) and leads to disparities in academic achievement during their school years; and in later life including lower earnings and family stability in adulthood (Hart & Risley, 1995, 1999; Heckman & Masterov, 2007; Rowe, 2008; Walker, Greenwood, Hart, & Carta, 1994). So, while the data and knowledge about the Word Gap and potential solutions for reducing the gap have been available for a number of years, to date, no coordinated national effort has yet been launched to address this important disparity.

The BWGRN was funded by the Health Resources and Services Administration (HRSA) within the US Department of Health and Human Services to reduce the number of young children who enter school with delays in early literacy and language (Bridging the Word Gap National Research Network, 2015). Among the several objectives of the BRGRN was the creation of a DDI to uniquely support the needs of intervention researchers and evaluators who are engaged in inquiry related to finding out "What works, For Whom? and Under what conditions? in bridging

the Word Gap. The DDI is intended to support the research agenda of the BWGRN (Bridging the Word Gap National Research Agenda, 2015) as well as other related intervention research.

The vison of the BWGRN-DDI is creation of a web-based knowledge system to support the work of researchers designing and conducting experiments. The intent is to increase the experiment rate and speed solutions for intervening with families, communities, and the nation to mitigate and prevent the Word Gap. The function of the DDI is to support researchers in learning and using the innovations of others, such that new inquiry builds on current knowledge, and research tools do not have to be reinvented over and over again. The motivation in using the DDI is achieving the leverage and advantage obtained by access to existing tools and knowledge that is not readily available in peer-reviewed publications or even a single database.

Because the DDI for intervention researchers does not yet exist, this report begins by sharing a vision of a DDI for language intervention researchers and research community. Remaining sections of the report describe the process used to determine the function and content of the DDI including a synthesis of focus group findings as reported to us by Network members who are engaged in intervention research. We conclude with next steps to move forward with this community of scientists in building their DDI.

GUIDING VISION FOR THE LANGUAGE INTERVENTION RESEARCH DDI

The BWGRN is committed to developing a DDI designed to support language interventionists and intervention research that ultimately prevents early delays in oral language from becoming deficits in early literacy and reading at the population level (Greenwood et al., in revision). **Similar to the DDIs previously discussed, we envision a networked digital system promoting the sharing and consumption of knowledge, data, tools, and technical research information uniquely supporting language intervention research.**

Its major functional components are (see Figure 1):

(a) **Creators** [Who provide tools, data and metadata[1]],

(b) **Discovery** [Locating results of a search],

(c) **Archive** [The base of information in the form of data, documents, and other digital assets], and

(d) **Users** [Who explore and access needed information].

As can be seen, this DDI supports searching, exploration, posting, accessing, and using data in various forms of research. It also is a system planned, supported, and contributed to by individual researchers. Thus, the DDI will support the language intervention research community of Users in their quest for and use of technical information needed to advance current knowledge. It will also enable building of greater capacity and ability to acquire knowledge for Bridging the Word Gap.

How will BWGRN Intervention Researchers Benefit from a DDI?

Intervention researchers need resources to plan and conduct next step experimental studies, including detailed information about interventions (e.g., video clips), measures, analytical methods, intervention effects sizes, and pathways. **This DDI will be an information system uniquely designed to inform researchers who design, propose, and conduct experimental studies of language promoting interventions at the child, community, and**

---

[1]Metadata = documentation describing the data or asset supporting searching and discovery by others

levels. *The value of such a system would be support and expansion of a distributed, diverse program of national research as well as the services and facilities necessary for research efforts to function at scale.*

This DDI will facilitate ongoing growth in assets and resources over time due to collaborative contributions from members of the community. With the creation of a DDI related to language intervention, we would expect to see increase in the number of studies focused on promoting early child language skills, greater collaboration among researchers, and more support for funding and resources aimed at interventions to help reduce the Word Gap. Over time, these intermediate outcomes would lead to a reduction in the word gap for many, improved early language outcomes (i.e., improved early vocabulary), and more children who are ready to enter kindergarten.

The language intervention research DDI will be an **online library** housing relevant information, tools, and data, including:

1. A large corpus of over 2000 peer-reviewed studies used for the BWGRN research syntheses (a Zotero database) for use in conducting future intervention research syntheses.

2. A library of evidence-based interventions. Information would be available describing the strength of existing evidence for each intervention

3. Information and discussion of new and recommended research designs and methods.

4. Online tools and resources related to the most recent advances in measurement.

5. Information and discussion of new technologies and information management tools relevant to BWGRN intervention research

6. Intervention (experimental) datasets available for download and use in secondary analyses.

Steps toward Developing the BWGRN Digital Data Infrastructure

To gather more specific information about these and other resources we reached out to the BWGRN community of intervention researchers for their suggestions using a series of small focus groups to evoke their needs and suggestions. **The purpose was to understand the needs of early language intervention researchers in greater detail, determine what is needed to support their work, and gauge their interest in contributing to and using the DDI.**

Research Questions

Considering the unique needs of BWGRN intervention researchers, we were interested in knowing the following from intervention researchers:

1. Do intervention researchers see a need for a DDI to advance their work?
2. What exactly should the DDI contain to make it useful?
3. How should the DDI be organized and operated?
4. How should users contribute to the DDI?

METHOD

Participants

To address these questions and provide guidance to a DDI development plan, BWGRN key staff (Co-PI, Charles Greenwood, and Project Coordinators, Anne Larson and Alana Schnitz)

met with BWGRN members (N = 18) known for their intervention research, we conducted 8 separate one-hour video conference calls, that took place June 22nd to July 6th, 2016 (see Table 1). Participants were asked to review a BWGRN-DDI vision statement document sent to them prior to the call. All participants reported having read or looked over the vision statement and none had questions about the document.

During each video conference call, focus group participants were asked to describe their primary areas of research and any current or recent intervention projects. Reported research designs of participants included randomized control trials, functional analyses, single case research designs, and longitudinal studies (also see Appendix A, pg. 19). They reported conducting research with low-income preschoolers as well as children with hearing loss, Autism Spectrum Disorders, intellectual disabilities, challenging behaviors, and language disorders. Researchers also reported work developing and evaluating strategies for parents to use in interactions with their children. Interventions were taking place in the context of home visits, in preschools, community settings, home-based child care settings, and clinics. Some researchers said they were interested in translational research to teach early childhood professionals about effective interventions to determine the best ways to scale-up use of effective interventions. Others were specifically interested in helping to move interventions into pediatric care settings.

Procedures

We used the focus group approach. We purposively limited the number of participants per session to no more than 4 to enable each to express their views and facilitate accurate recording of responses. Following participants' descriptions of their intervention work, all were asked to respond to our questions about the BWGRN-DDI (see Appendix A).

One BWGRN staff member facilitated the conversation and another recorded responses live during each video conference. The collective set of raw narrative notes and conversational comments was synthesized to topically distinct themes related to needs and content addressing needs of the intervention researchers.

RESULTS

Research Needs and Content

Overall, participants provided 136 comments related to 9 themes, and additionally points afterwards via email (Appendix A). Most participants indicated that a DDI was a new idea for them but that the vision statement gave them new insights regard how useful such a system could be in supporting experienced and emerging researchers. Respondents mentioned that the DDI could provide more useable information related to research reports from both current and past research projects, and could help inform future research. Some participants suggested a need for resources that could help bridge research and practice, and others were focused on sharing and improving existing research methodology in terms of more detailed information about specific interventions, and measurement tools used within intervention research.

The following section summarizes responses from focus group participants when they were asked about the resources they need for conducting intervention research, as well as their responses when asked about the content that would be useful to have in the BWGRN-DDI.

Intervention Library. Focus group participants suggested having a place that describes current interventions and the research that is tied to those interventions. Mentioned repeatedly across conference sessions, different participants suggested that it would be helpful for the DDI to have a repository of language interventions and specific information about how to implement them. Many thought that using a standard template to describe the interventions would be helpful so that descriptions would include information that could easily be compared across

interventions. Another suggestion was to include video clips of the intervention to illustrate delivery and implementation.

One participant reported difficulty working on a meta-analysis because the written descriptions of interventions did not provide enough information. Suggestions for intervention study details to be included were: information about the specific populations that have used the intervention or that would be recommended for receipt of the intervention, information about dosage that is needed for children with specific needs, and protocols used by those carrying out the intervention. Overall, focus group participants suggested providing more detailed information on interventions than can be gathered from the literature (i.e., specifics about the training and coaching needed to deliver the interventions) would be useful. In the focus group with community-based researchers in particular, resources specific to running community-wide interventions would be desirable.

Measures Library. Several focus group participants said they would benefit from knowing about and having access to measures (published and researcher-created) being used in language intervention research. The information would help researchers in selecting measures when designing studies (based on design, research questions, participant characteristics, etc.). They also suggested that it would be useful to have resources describing the procedures for using a measure, (i.e., codebooks and reliability procedures for observational measures). One participant suggested that the DDI measurement section should be smart enough to allow for a student to pose a research question and search for measures likely to help them answer it. Others suggested that by having a common place where measures used across studies could be identified and described, researchers would be able to compare how different measures have worked for different populations and how measures compare to one another (e.g., how does LENA compare to the PLS-5, for example) – possibly leading to adaptations in measures or development of new measures.

Other measurement suggestions included accessing shared codebooks and reliability procedures for coding child and adult linguistic interactions, as well as shared fidelity measures for particular interventions. These resources would allow for interventions to be compared more systematically across studies because of similar measurement. Such resources may also help researchers develop a group of commonly shared codes. One participant specifically suggested listing the contact information of people well-versed in a particular coding procedure so that they could be contacted regarding specific questions. Beyond language outcome measures, other participants recommended describing tools that look at constructs such as parent attachment, child and parent interactions, child observational codes, and satisfaction with an intervention.

Analysis Resources. Across sessions and participants, it was stated that language intervention researchers need more effective ways for analyzing data gathered within and across studies. It was stated that intervention researchers need to know how an intervention is working for individuals and/or for participants with specific characteristics, and factors that contribute to participant attrition. Focus group members suggested working together to learn more about matching studies and determining the best ways to build comparison groups that might serve as counterfactual controls. Participants also suggested the need to develop new analysis strategies for integrating data across studies to increase power, for learning more about minority populations, and for better understanding intervention effectiveness and potential generalizability.

Data Resources. The topic of data sharing came up in each conference call. Focus group participants said it would be helpful to provide access to data from various intervention studies (including datasets from single sessions) for reanalysis. In particular, it was suggested

that having more transcribed and labeled LENA data that could be used for comparison to other populations and settings would be useful. More specifics on data sharing are provided below.

In addition to data sharing within the BWGRN DDI, participants also identified resources they needed or about data sharing arrangements being used in other fields of research. Some participants suggested that the DDI house information about how to share data with organizations that require doing so (or soon will), how to meet HIPAA standards for data storage and plan for updates, information about the benefits/disadvantages of sharing data along with a journal submission (if available), and how to save data so that it can be used/ updated most easily in the future (i.e., what is the best way to save LENA data so that it can be reanalyzed with any new algorithms that come up in the future?).

Several researchers reported using digital tools, including RedCap, for example, a web-based program for surveys and overall study data management (https://projectredcap.org/). Although some participants pointed out that RedCap is not a perfect match for psychology-based research, it is a system that people are trying to learn that needs support. Some participants suggested their willingness to share templates of coding protocols that are based on fine-grained, individual item responses on standardized assessments saved in RedCap. Others asked for specific resources such as a procedural manual to be shared (i.e., how to run reports) so that everyone who is just starting out with RedCap will be able to fully use the program.

Video Library. Several focus group participants were interested in the idea of having high-quality videos to examine real-life interactions. Researchers want access to videos to develop coding schemes, practice coding, test reliability on coding, and compare the coding systems that are used across studies. There were concerns about gathering permission to share videos on the site. Suggestions for dealing with permissions included having families agree to participate as samples so that they know going into a video-taped session that the videos will be widely shared. Researchers also suggested gathering information from technology professionals who may know more about procedures for de-identification. There was also recognition that we should not duplicate video available in other sites (i.e., Databrary).

Research Articles and Synthesis Library. Focus group participants were interested in accessing the intervention research literature contained in the BWGRN electronic Zotero Group Library – this was noted to be especially important for community researchers who may not be able to easily access up-to-date research articles. Suggestions were made to develop a system for bringing in new research articles to the Library as well as a process for reviewing articles so that they be added to the database in a systematic way. In addition to requesting access to the BWGRN Zotero database and workgroup syntheses, focus group participants suggested that the DDI house research-based, synthesized information about specific language-promoting practices. These supplementary syntheses could identify practices as "proven" or "promising" and link readers to supporting evidence. Another function of the practice-specific syntheses would be to make suggestions about future research needs.

On Contributing to the DDI: Barriers to Data Sharing and Possible Solutions

Most interesting was the willingness of participants to share part of their work. Many of the focus group participants indicated that they thought members of the intervention research community would contribute to the DDI if technical procedures and support requirements were clearly worked out and easy to use. Additionally, focus group members indicated that they were not exactly sure how they might contribute to the DDI but most stated they would be willing to help.

Researchers reported possible barriers to sharing their data with others in the DDI. One concern was the fear that others might publish findings before the owner, thus jeopardizing the professional advancement of the owner. Another was simply knowing what others actually

would do with the data. A different concern was not having time to organize the data in a way that might be needed for the site. A suggestion of one researcher who was reluctant of posting data for personal reasons, suggested that a solution might be to house descriptions of specific data sets in the DDI so that the owner could decide individually, who they might share data with and who they might consider for collaboration. Others suggested careful vetting of potential users of the DDI to prevent misuse, and to establish a gateway of trust and common mission. A possible solution here would require that potential users are required to post their own profile on the site as part of their application to become a user such that site managers and active community members are able to confirm each applicant's credentials and understand their legitimacy to access the system, give their resume, research interests, and current research.

Last was uncertainty of the IRB procedures that would be needed to allow sharing of the data when blanket approval for data sharing was not currently part of their IRB process for either current or past research project, thus a barrier to sharing data. Other researchers brought up concerns about privacy protection laws in specific states (i.e., Maryland) related to LENA, as well as other recorded data. One solution proposed was to limit data sharing to only de-identified data. Another solution was to limit data sharing to only future projects where permission for data share was obtained prospectively.

There were several practical questions regarding data upload procedures and how it might be organized on the site. A common concern was that standard requirements be clear in advance so contributing researchers will not be burdened in preparing their data for posting (e.g., IRB and the file format would be required, etc.). As noted above, in lieu of posting data, a brief description of the data following a consistent format for all entries, could be posted with contact information for those potential users who might want to follow-up or collaborate.

How Should the DDI be Organized and Operated?

Need for Consistency. Many intervention researchers called for the DDI to bring consistency in the methodology used in studies promoting language development. For example, several focus group participants suggested the DDI create a standard protocol for describing demographics (i.e., SES, home language), measures, and interventions to provide consistency across studies. Additional factors for the protocol include reliability standards, and intervention fidelity measurement. Focus group attendees agreed that, as a field with greater consistency and transparency we would be able to move the science forward more quickly.

Site Structure and Organization. Most participants agreed that the idea of having an open-source repository was a good way to start the DDI. One participant suggested that the infrastructure needs to "create momentum" – that it needs to be accessible and attractive to new researchers who would initiate its building and continue its development.

Several people recommended that the DDI platform be easy to use and sustained by individuals who would provide technical support for persons trying to give information so that they wouldn't get frustrated and give up. Others suggested that the DDI be monitored by at least one primary 'gatekeeper' who would track users, assist with soliciting information to post, vet incoming resources before posting, and continuously organize content.

A few participants thought that the proposed networked DDI would potentially attract people to visit the site and keep them coming back to visit if it proved to be useful and not burdensome. The DDI they thought would allow users to interact not only with the content of intervention research but potentially one another as well.

In addition to creating connections within the DDI, others suggested that the BWGRN connect the DDI resource with its other communication tools including the website, *The Bridge* e-

newsletter to members, and social media so people who might not visit the site on a regular basis will do so given learning of these new and useful reasons to do so.

Suggestions were also made to incentivize contribution to the site. For example, one participant suggested having people who donate to the site gain free access to measurement tools created by Network members or obtain a special status such as a "BWGRN-DDI Investigator" that could be used on a CV. Another suggested that the DDI might ask applicants to contribute data or a measure, or video clip, etc., to the infrastructure as a token toward access. Such contributions could be based on a hierarchy so that each applicant's level of experience be considered so that the contribution is reasonable. For example, doctoral students might be asked to code studies, whereas, a BWGRN Emerging Scholar might be asked to help with site management, versus a seasoned researcher or current BWGRN member who would be asked to contribute a RedCap procedure for coding videos of parent-child interaction or similar asset to the site.

SUMMARY

Overall, focus group participants were interested in having the BWGRN DDI also serve as a source for professional development, teaching, and as a way to connect researchers to resources. One participant remarked that this project had the potential of being more accessed than other datasets currently used within the field (i.e., National Early Intervention Longitudinal Study data), suggesting that it will be more active, versatile, and usable. Many BWGRN researchers who joined the video conferences were excited to have a repository to pass on to the next generation of researchers, to think about new research questions to address the needs of children and families to help the field progress. Participants are eager to see what comes out of this work and are willing to help anyway they can.

NEXT STEPS

Based on this input and lessons learned from the intervention research community and the earlier background discussion, we propose to undertake these next steps toward development of a DDI.

Step 1. Prioritize the content to be developed on the site based on this input from the intervention research community previously described and reach out to community members for access to their content. We will start with useful content that can be posted with only permission of the developer. We will take on other content that requires special permissions later.

Step 2. Develop site management procedures, policies, and map out and implement the processes users will need to use to obtain credentials to access and use the site.

Step 3. Train managers as needed to supervise and monitor site utilization going forward as needed.

Step 4. Design and develop a prototype website capable of functioning as a networked DDI that can be accessed on its own on the Internet or through a link placed in the BWGRN website for easy access.

Step 5. Develop the Home page of the site to include the major content libraries and navigation within and across libraries (as described above). We will program the navigation and test and improve function.

Step 6. Develop the basic content within each library, obtain the assets that are the content, link to navigation and test and improve function

Step 7. Market the DDI broadly to BWGRN and other intervention researchers using the website, newsletter, social media, presentations at professional conferences, and word of mouth.

Step 8. Seek external funding to fully develop the DDI based on the data and experiences of using the prototype with a goal of making the DDI an enterprise level website capable of serving more researchers with national and international reach.

CONCLUSION

Like historical developments in science generally in the past 20 years, intervention research to Bridge the Word Gap has amassed an advanced technology and base of knowledge regarding interventions that work, methods for experimental research for different purposes and degree of complexity, capacity for "big data" and digital applications, and the potential for prevention at the population level in future. The need for a DDI for the BWGRN community is clear, and the inquiry reported in this document has set the stage of its development, use, and refinement going forward in the next several years.

References

Bell, G., Hey, T., & Szalay, A. (2009). Beyond the data deluge. *Science, 323*, 1297-1298.

Bridging the Word Gap National Research Agenda. (2015). *National Research Priorities for Bridging the Word Gap*. Retrieved from Kansas City, KS: http://www.bwgresnet.res.ku.edu/wp-content/uploads/2016/01/BWG-Research-Agenda-12-15-15-Final.pdf

Bridging the Word Gap National Research Network. (2015). Home page.   Retrieved from http://www.bwgresnet.res.ku.edu/

Carta, J. J., Greenwood, C. R., Walker, D., & Larson, A. (2016, February). *A national survey to identify the important priorities for bridging the word gap.* Paper presented at the Conference on Research Innovation in Early Intervention, San Diego, CA.

Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science, 16*(2), 234-248.

Greenwood , C. R., Carta, J. J., Walker, D., Gilkerson, J., Watson-Thompson, J., Larson, A. L., & Schnitz, A. (in revision). Conceptualizing a public health prevention-intervention framework for bridging the 30 million word gap. *Clinical Child and Family Psychology Review*.

Greenwood, C. R., Schnitz, A., Irvin, D., Tsai, S. F., Crutchfield, S., & and the Bridging the Word Gap National Research Network. (Submitted). Automated Language Environment Analysis (LENA): A Research Synthesis. *American Journal of Speech-Language Pathology*.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore: Brookes.

Hart, B., & Risley, T. R. (1999). *The social world of children learning to talk*. Baltimore, MD: Brookes.

Heckman, J. J., & Masterov, D. V. (2007). The productivity argument for investing in young children. *Review of Agricultural Economics, 29*(3), 446-493.

Radesky, J. S., Carta, J., & Bair-Merritt, M. (2016). The 30 Million–Word Gap: Relevance for Pediatrics. *JAMA Pediatrics*. doi:10.1001/jamapediatrics.2016.1486

Richards, J. A., Gilkerson, J., Paul, T., & Xu, D. (2008). The LENATM automatic vocalization assessment. *LTR-08-1*. Retrieved from https://www.lenafoundation.org/wp-content/uploads/2014/10/LTR-08-1_Automatic_Vocalization_Assessment.pdf

Rowe, M. L. (2008). Child-directed speech: Relation to socioeconomic status, knowledge of child development and child vocabulary skill. *Journal of Child Language, 35*(1), 185-205.

Thanos, C. (2011). Global research data infrastructures: The GRDI2020 vision. *GRDI2020 Roadmap Short Version*, 25. Retrieved from http://www.grdi2020.eu/repository/filescaricati/fc14b1f7-b8a3-41f8-9e1e-fd803d28ba76.pdf

Walker, D., Greenwood, C. R., Hart, B., & Carta, J. J. (1994). Prediction of school outcomes based on early language production and socioeconomic factors. *Child Development, 65*, 606-621.

Warren, S. F. (2015). Right from birth. Retrieved from https://www.lenafoundation.org/wp-content/uploads/2015/08/RightFromBirth_Warren_5.12.2015_v.3.pdf

Xu, D., Yapanel, U., & Gray, S. (2009). *Reliability of the LENA^TM Language Environment Analysis System in young children's natural home environment*. Retrieved from

Table 1

Focus Group Call Participants and Meeting Schedule

| Participants	Date
1. Jay Buzhardt, Ilene Schwartz, Dale Walker	June 22, 2016
2. Kathy Bigelow, Megan Roberts	June 22, 2016
3. Megan Bair-Merritt, Susan Sandall	June 27, 2016
4. Erin Barton, Judith Carta	June 27, 2016
5. Jill Gilkerson, Courtney Hawkins	June 29, 2016
6. Steve Warren, Carol Trivette	June 29, 2016
7. Sara Johnson, Juliann Woods, Howard Goldstein, Ann Kaiser	July 1, 2016
8. Dana Suskind	July 6, 2016
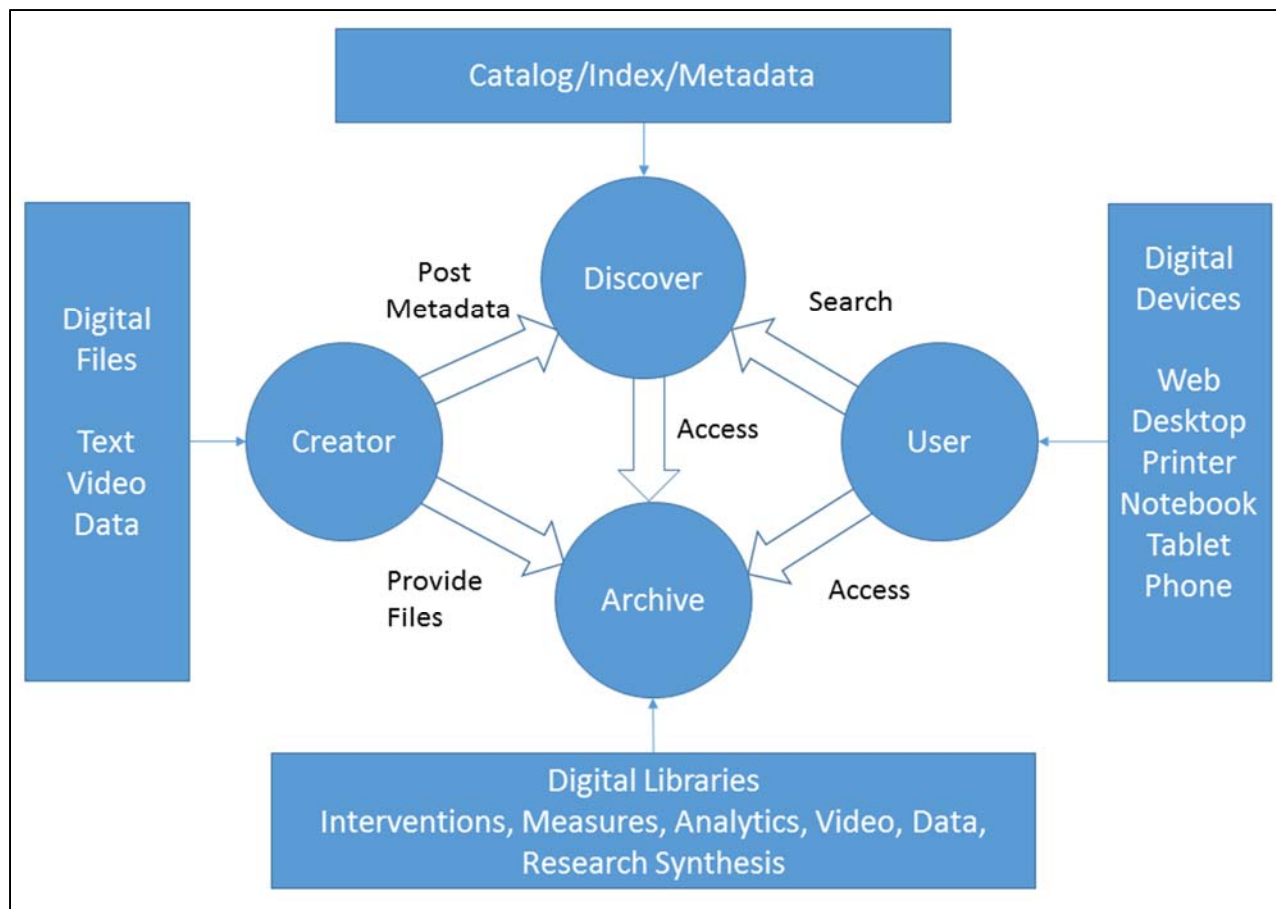
Figure 1. *Illustration of a BWGRN data infrastructure (with a Web Emphasis).*
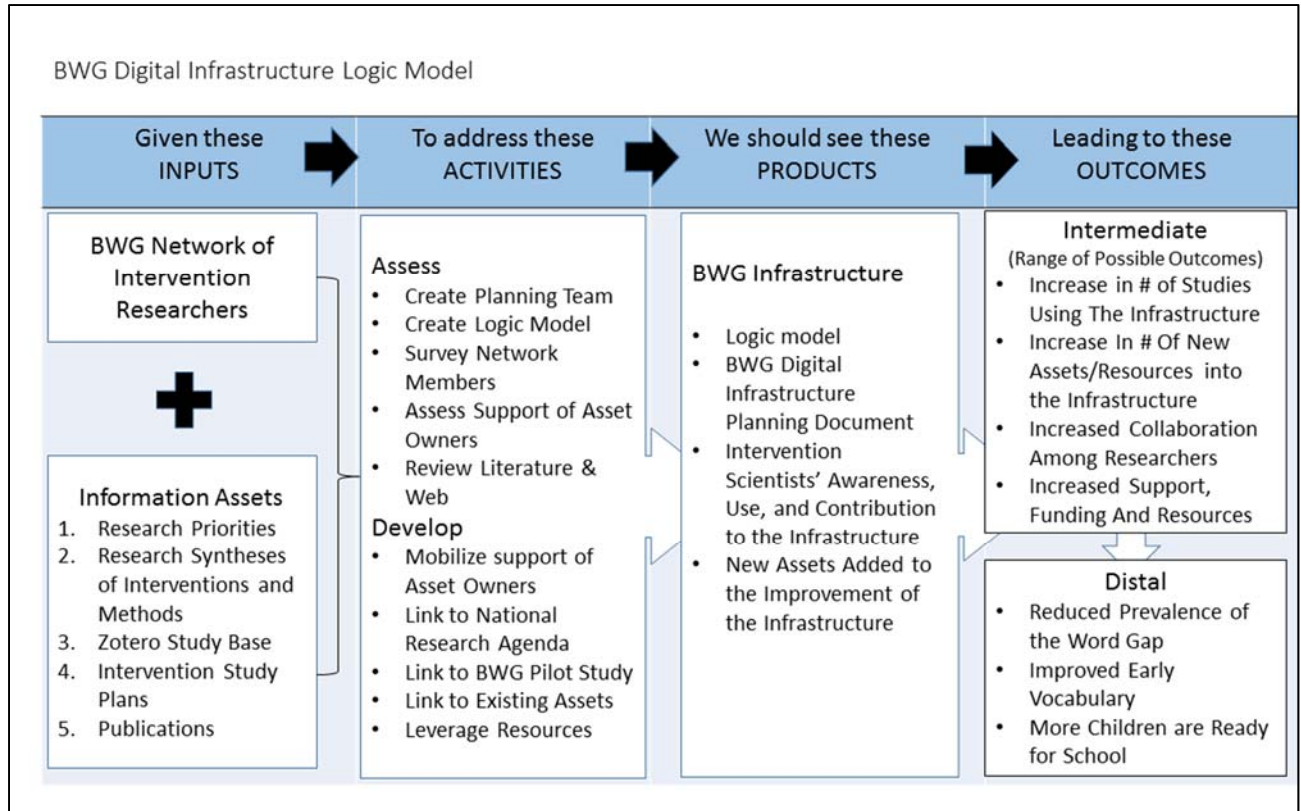
Figure 2. BWGRN DDI logic model.

Appendix A

The Unique Nature of BWGRN Member's Reported Intervention Research

We thought it was important to document the research themes reflected in this group of language intervention researchers that differentiate their needs from noninterventionists. What follows is a sketch research themes engaged in by these participants.

Many of the members were currently running randomized trial studies of interventions, some in in schools, and one parent-implemented at home, one mixed with kids with ASD in clinic/home intervention, one with kids with Down Syndrome in home/clinic intervention. Others involved children with language delays and Spanish-speaking parents.

Several single case experimental study were underway as well. All were collection video data as well as standardized assessments. Others reported intervention development projects in the IES adding new features to existing tier 2 curriculum to explore table applications and extend into home as well as preK practice. Studies of the effects of coaching interventions carried out by community providers in homes for kids with severe disabilities was reported. Also doing some things with technology. Some other members were conducting biobehavioral studies looking at how poverty effects physiological aspects of the brain.

A number of group members mention interest in translational research – translating things that work for professional development so that they can be used by a large audience. Recent doc students have been conducting intervention research with a focus on very reasonable, focused interventions (e.g., talk during snack time)

There was a primary interest in seeing how evidence-based interventions can be moved into pediatric care, teacher training on implementing EBP, and parent training to implement social-emotional practices in the home. There also was an interest in home visit models as well as use by early childhood educators in child care, center- and home-based. One member reported her focus on work in state systems with respect to promoting the uptake of interventions by state providers using an implementation frameworks approach.

In commenting on the value of an online data infrastructure for intervention research, one member had built a web platform for their 5 year longitudinal study (Spent about $250-350,000) to keep track of participants and data (200 participants and 17 data points for each person). The program houses their curriculum, gives feedback to home visitors and can be used for analysis. With this system in place, when home visitor completes a fidelity of implementation assessment, results go straight into the infrastructure. This program is housed at the University and the initial platform was designed by external consultants.

With respect to digital information and tools that are highly relevant to their work, several members reported using uses the clinical trials registration (www.clinicaltrials.gov), others reported Redcap – as a way to manage projects and also as a way to collect data. *REDCap (**Research Electronic Data Capture**) is a browser-based, metadata-driven electronic data capture software solution and workflow methodology for designing clinical and translational research databases. It is widely used in the academic research community: the REDCap Consortium is a collaborative, international network of more than 2000 institutional partners in over 100 countries, with more than 400,000 total end-users employing the software for more than 200,000 ongoing research studies* (online at Wikipedia), However it was noted that the existing RedCap Redcap modules are not geared towards studies that have a psychological component, but are more of a biomed focus.

Summary. In this summary of the reported research activity of intervention researchers, we learned as expected that they are engaged in conducting experiments and in translating

successful interventions from research to practice, through professional development. They are using measurement batteries that include a range of diverse measures of adult and child behavior and communication including fidelity of intervention. Taken collective, the unique needs of this community of researchers is clearly demonstrated and largely different from that of others who do not conduct experiments or engage in training and professional development design to improve implementation quality and children's outcomes.

Members' Responses to Focus Group Questions

1. What needs do you have that could be supported by a data infrastructure?

1. We need a better way (not RCTs and not SCRDs) to gather information about how interventions are working. Can we combine research to get more power?

2. Wants to learn from the differences on individual responders

3. Ways to be able to compare data on assessments that are completed in different languages

4. Putting up the lit review (Zotero database) would be helpful

5. Having something related to measures would be useful. A doc student has a specific question, send them to the database to look at measures that can help answer the research question

6. Would also be nice to have a list of people who are well-versed in the coding procedure (i.e., through CSBS).

7. If we had a standard protocol to be able to reach a consensus on SES, home language and literacy environment (at least among people doing BWGRNresearch). Language sample coding is more difficult.

8. Would be helpful to have resources for people who are working on matching studies (what n do you need). How do you run reports in RedCap? Feel like everyone is rediscovering potential contributions of RedCap rather than sharing the information.

9. Every time they need something, they just reach out to the individual. Likes the idea of a discussion board though.

2. What content domains would have value to you in the data infrastructure (e.g. measures library, literature library, intervention library, etc.)? What exactly should the Infrastructure contain? What would interest you enough to use the infrastructure in your work?

30. Would be great to have a methodology-based infrastructure (what measures are other people using? how often are people measuring? Can we make comparisons across studies more easily?)

31. Get some consistency across methodology

32. See what people have done, what results they had

33. Idea to have video clips of what the intervention looks like

34. Descriptions of interventions tied to research that has been completed on them

35. Place to hold packaged interventions (split by particular intervention (i.e., dialogic reading) and specific population that the intervention has been used with/supported for

36. Future research needed section – place for students/emerging scholars to go and quickly identify what the next steps are in moving forward intervention to close the word

gap (this may be a section within the intervention library --- what exists? what is it? what is needed next?).

37. Limited use of video clips is attractive, more expanded use seems like it would be difficult with permissions.

38. Having people report interventions along with data or publications

39. Would be helpful to have some consistency in what an intervention is called and what it looks like.

40. Would be helpful to understand the people who are accessing the system, what they are working on.

41. Data sets from single sessions would also be helpful to house in the data infrastructure for reanalysis

42. Judy: Good video files (high-quality videos and high-quality interactions)

43. After syntheses are completed with BWGRNit would be nice to have syntheses for different language-promoting practices. Create a searchable website so that a user could identify which practices are proven vs promising, or those which don't have enough data. Look across studies using that intervention (what participants has it been affected for, what does the fidelity checklist look like, what kind of dosage has it been implemented with?). Purpose would be for someone to identify a practice for programs/practitioners and research (where do we need to go next)

44. And policy – children with these needs need this dosage of this intervention. Would also be helpful to have information about the training and coaching needed to delivery these interventions

45. Some of this might overlap with DEC recommended practices

46. Having a place where community researchers can access research articles

47. Intervention protocols so that people can replicate approaches, and ask contributors to uphold simple standards when uploading contributions

48. Observation codes from intervention studies would also be nice to compare, have access to

49. Film clips of actual intervention sessions to see what things actually look like

50. Working on a meta-analysis, and they really don't know what people did in different interventions – access to datasets, protocols, etc.

51. Would be great to have things to pass on to the next generation – think about what the next level is, where do we need to go as a field?

52. Interested in this database as a source for professional development

53. Loves the idea of video to be able to examine real-life

54. List of journals, what they accept, what they cover, links to their author information

55. Fidelity assessment examples (e.g., they have a fidelity measure for the SPA – play-assessment)

56. Thinks RedCap would probably work with BWGRNon things that are important to behavioral researchers

57. Information about how to share data since we will all need to do this (new requirement from IES)

58. Saving LENA data – if you compress the data will you lose the ability to code the data with new LENA measures that come up in the future.

59. Put together prospective IRB language for use of data on this website.

60. Can we be helpful in developing HIPPAA strategies for sharing data

61. Have to budget for updating HIPPAA information so that you can update

62. Have some recommendations for SES and then people can say "we used the BWGRNprotocol for SES"

63. Have a template for reporting parts of the intervention, including dosage.

64. Clarifying that we are thinking about this as an interactive platform

3. Would you be willing to post your content (e.g., measures, literature, intervention, datasets) to the infrastructure to inform others in their work?

65. Thinking about submitting a brief report along with data that includes key words to search.

66. We would need to look into how far the Zotero database can be shared with copyright limitations on research articles

67. People are always afraid to put their data up

68. Would be willing to put all measures for what she has in Redcap (i.e., data dictionary for PLS-5 items and response options)

69. Need to have specific details about how to make a contribution to the infrastructure

70. Suggestion for two phases, BWGRN gathers 3 videos then asks 5 labs across the country to code them. Eventually have all interventions submit their codes and codebooks.

71. Pretty protective of her data but could imagine posting information about the data set in order to build collaborations. For example saying something like, I have this data set which contains this information and if someone is interested in collaborating they can contact me.

72. Hesitant to post the app but willing to post information about it to promote collaborating with other populations

73. Would really like to post things (i.e., LENA data)

74. People would contribute lessons-learned

75. People are not going to give you the data until they are done working on it, and even after that they will probably only give you data if it is required.

76. Willing to share lots of things that are being developed but would like them to be vetted/discussed to make sure that they have something that has broad applicability.

77. Have families that agree to participate as samples so that the families know going into it that the video will be widely shared.

78. Could be even more difficult to collect LENA data given differences in State laws on IRB.

79. Would be helpful to have standard ways of describing demographics and measures that we have for a particular data set. For example, we know that only 50% of WG studies reported SES!

80. Would be nice to have people share their data dictionaries

81. Perhaps posting the SPEAK which is still in development. Using executive functioning from Susan Landry and Anne Fernald's *Looking is Listening*.

82. It is important to have a community/network, this will keep people coming back

83. ECRQ is now accepting all data to be submitted along with manuscript. If paper is published they will publish the data and a couple of paragraphs about it the data separately. Program is called: "Data in Brief" (provides a way for researchers to easily share and reuse data – this is likely the journal in itself)

84. Can we have a list of things we need that people have to do (choose one from the list) in order to gain access to the infrastructure?

85. Give the people who have contributed something an "investigator-type" status so that they can put it on their CV. Down the road, there could be different levels of contribution.

86. Eager to see what comes out of it and would love to help out how she can.

87. I am great fan of RedCap for data management. We use it for everything. We have developed templates for managing single case design studies and group design studies that include everything from screening, to video links for coding, to fidelity, and of course outcome measures exportable in multiple statistical formats. Happy to have someone from my team talk about that if that is useful or at some point to have the RedCap developers who are here at VU do some presentations.

88. Measures... yes, absolutely, with precise variable definitions, formula for calculations of actual values reported, access to coding protocols and minimally published assessments (referenced but not available from a publisher).

89. It would be outstanding if the new BWGRN could sponsor a series of mini-conferences or web-based work groups on some of these topics.... with conference presentations linking to the online resources.

4. How should the DDI operate in your view?

90. Not sure how we would contribute

91. How would data be uploaded? A list of files from individual researchers? A more structured database?

92. Need to look into permission to release video at this level

93. Organize message boards to post questions or responses. Start with a group of researchers who are willing to start (e.g., LENA – what it is, why it was used, how was data collected?)

94. Provide an opportunity to build a community and allow people to interact

95. Connect this resource with a newsletter that shares what is new

96. Input system needs to be very user-friendly and streamlined in order to want to share.

97. As long as something is reciprocal, it seems like it would useful

98. Needs to be someone looking at what is coming in based on predetermined quality and content.

99. Have it be an open-source repository so that it is of use to multiple researchers but ALSO foster collaboration

100. Think of something for a couple of different user groups to get it going. Would need somebody to manage everything (making sure things are accessible, people are using it). Start with someone who is part of development to start creating it

101. A little worried about the permissions that parents have already signed, think it might be easiest to share after

102. Could have people who donate data get free access to the LENA Snapshot

103. Needs to be easy!

104. IRB issues need to be addressed up front and let people know specifics about what can be submitted/how to submit it (protocol that says what contributors need to do, here is how the data would be used, here is when you would need to give additional information, describes our expectations, etc.). Also need to think about what format we would want the data in – what standards are there? Can the standards be as broad as possible so people don't have to adjust too much.

105. Would be willing to help.

106. Would be very open for collaboration around the data that will come out of the longitudinal study (doesn't think they would be able to share their actual data on the site but would like to collaborate – only so much one person can do with all the data!) Would be interesting to hear from each member of the Network to see if they can think of one thing that they could contribute to the infrastructure.

107. Would there be a gatekeeper?

108. Yes, someone to decide who can join and what can be posted

109. Organizing the infrastructure. Could this be something people can cite? Would each intervention have the same content? e.g., video, description, materials, research articles associated with that intervention, etc.

110. Who is the target audience? Researchers *and practitioners* looking for interventions?

111. Autism Speaks has a video glossary that is useful for teaching about different interventions. Not sure how something like this would be used for research but it would be fabulous for teaching.

112. Reminder not to focus on everything or duplicate anything that already exists

113. The infrastructure needs to create momentum, needs to be accessible and attractive to doc students and new researchers (these are the people that will build it and keep it going). Could students access this for their early research projects? Could have mini-courses.

114. Think about using the Emerging Research Scholars to help make a list of what the applications through the infrastructure might be.

115. Consider using Slack or an existing platform

116. We need to figure out a way to get the word out that the data infrastructure is there and know how to access it

117. This project has the potential of being more accessed than other data sets (e.g., NEILS data). This infrastructure sounds much more active, versatile, and usable compared to a large dataset

118. We don't want to duplicate anything in this infrastructure that is in another (e.g., describe NEILS in our infrastructure with link to their system)

119. Who will handle the building/interface? That will be really important for usability/accessibility

120. NSF funds data infrastructure projects, would like to look into this for the future. Likely we will have several stages of implementation – might begin with links from our website, then move to website that can be accessed through usename/password, and finally a nice looking/easy-to-use version of the password-protected website

121. Where do we begin to develop something that will be helpful to the field? Idea to prioritize having a video library that can be coded in different ways (even for programs that want to be certified in certain kinds of practices).

122. Calling this a "data infrastructure" may be confusing to some people because it is more than just a compilation of raw or metadata. Thinking about the name will be important – make sure it describes what is included.

123. How do we incentivize people to contribute?

124. Are you thinking about doing this in pieces? Possibly start with small group of people who have things they would like to put up on the site (based on feedback from focus groups), start with prototype, and invite people to become registered users.

125. Community-level interventions could have a different sort of resources (i.e., how do you get the mayor involved, how do you gather people together?)

126. The questions they always get are "can we look at your curriculum?" and "how do you do it?" (e.g., how frequently are you meeting families, how much do you pay families, how do you get funding)?

127. Idea to add in information about funding opportunities and a place where people can look for collaborations so that they can apply together

128. Responding to what she thinks about slack…if you have someone curating the site and putting new content on will be good. Self-sustaining will be hard to get people back.

129. Emerging research scholars could have some responsibility for adding/editing the site

130. Need to work out access issues to deidentified data. Not necessarily the raw data but the next level that is summarized and can be used for statistical programs. Not necessarily important for replication but work out ways for people to be able to combine data sets (modifying variables, collapsing variables, etc.). Think about how to deal with IRB approval, permission for people to be able to use these datasets.

131. Would it be possible to access technology professionals who know more about how to make video usable (i.e., blurring out faces).

132. Start small with items that don't require special permission.

133. Have someone review new articles using a pre-developed protocol to see how to add them to the Zotero database

134. We should focus least on the data portion of this website because it is going to be the most challenging to start up and the least known for what will happen with the information in the future.

135. Wondering what the timeline and next steps are.

136. Clarifying that we will end this funding period with a plan about what will go into the data infrastructure, then move to a prototype with people who are interested in contributing. Ask for funding, see if it can be almost self-maintained with assistance from graduate students/early researchers.