نشریه الکترونیکی سازمان کتابخانه ها موزه ها و مرکز اسناد آستان قدس رضوی
دوره یک شماره ۵
بهمن و اسفند ۱۳۸۸

بازیابی یکپارچه داده ها و انتشارات پژوهشی در کتابخانه های دیجیتالی[1]

مولفان: ماکسیمیلیان ا ستمفابر؛ بنجامین زاپیلکو
مترجم:حسین خسروی[2]

چکیده

در حال حاضر کتابخانه های دیجیتالی با چالش بزرگی برای یکپارچه سازی انواع قالبهای مختلف اطلاعات پژوهشی (مانند انتشارات، داده های اولیه، پرونده های شخصی، پرونده های سازمانی، اطلاعات طرح ها و غیره) روبه رو می باشند که برای این مشکل تا این تاریخ هیچ مدل عامی جهت سازمان دهی و بازیابی وجود ندارد. این مساله مشکل ساختاری و عدم تجانس معنایی را سبب می شود و باعث پیدایش طیف گسترده ای از استانداردهای فرا داده ای، واژگان و روشهای نمایه سازی که برای انواع مختلف اطلاعات استفاده می شود، می گردد. در این پژوهش، نه تنها بر یکپارچه سازی داده های مرجع برای انتشارات و داده های پژوهشی در علوم اجتماعی متمرکز شده بلکه نتایج آن برای رفع مسائل موجود در حوزه های دیگر نیز به کار گرفته می شود. ما مدلی برای بازیابی یکپارچه داده های واقعی و متنی که روشهای سنتی نمایه سازی محتوای انتشارات را با روشهای جدید ترکیب می کند، ارائه می کنیم. اما به ندرت این روش ها بر مبنای رویه های هستی شناختی هستند. روشهائی که برای ارائه اطلاعات پیچیده مانند آنچه که در محتوای داده های پژوهشی بکار می رود، مناسب تر است . مزایای استفاده از این مدل عبارت است از: ۱- سادگی استفاده مجدد از نظامهای سازماندهی دانش موجود ۲ - کاستن از مجموعه فعالیتهای موردنیاز جهت الگوسازی حوزه ای با استفاده از هستی شناسی

کلید واژه های موضوعی: بازیابی اطلاعات، کتابخانه های دیجیتالی، یکپارچه سازی اطلاعات، داده های پژوهشی، هستی شناسی، سازماندهی دانش، معماری اطلاعات .

۱- مقدمه :

درطول چند سال گذشته، کتابخانه های دیجیتالی بر اساس نقش و کارکردشان تغییرات گسترده ای برای کاربران متخصص خود داشته اند. (DELOS, 2005) نتایج چند پژوهش (POLL, 2004) نشان می دهد که بهره برداری یا پیوند دادن به فراداده ها از منابع مختلف و قابل دسترس ساختن آنها برای بازیابی با به کار بردن فقط حداقل استاندارد های فنی مرتبط با ویژگیهای داده ای و بازیابی برای پاسخگویی به نیازهای اطلاعاتی کاربران دیگر کافی نیست. کاربران توقع زیادی دارند تا نوعی یکپارچه گی قوی در انواع منابع اطلاعاتی (اعم از منابع تمام متن، منابع کتابشناختی، پژوهشها و سایر اطلاعات اولیه، داده های رشته های زمانی، اطلاعات طرح ها، پرونده های محققان و غیره) پدید آید. این تمایل نشان دهنده استفاده آنها از این نوع اطلاعات در مراحل مختلف و در ترکیبات مختلف در سراسر چرخه تحقیقات است. به ویژه در علوم اجتماعی، جایی که از یک سو آرشیو داده هاست و داده های تجربی را که با جزئیات بسیار در سطح بین المللی سازماندهی شده اند مستند می کند و شناسه های مدخل اختصاصی برای موجودی آنها ایجاد می کند و از سوی دیگر این اطلاعات و زیر ساختها به منابع موجود در کتابخانه ها و مراکز اطلاعاتی کمترین پیوند ممکن را دارند. این موضوع نه تنها همکاری فراهم کنندگان اطلاعات را در سازماندهی همکاری خود در گردآوری همه منابع در کنار یکدیگر به چالش می کشاند بلکه سبب پیدایش سوالات پژوهشی درباره چگونگی یکپارچه کردن اطلاعات پژوهشی به لحاظ فنی، ساختاری و معنایی نیز می شود.

پشتیبانی از چرخه کامل زندگی داده ها از جمله مدارک همراه یعنی نسخه های مختلف پرسشنامه ها، مجموعه داده های نهایی یک پژوهش، دستورالعمل نامه ها، نمونه های توزیع فراوانی و خلاصه آماری برای

---

[1] Stempfhuber, Maximilian; Zapilko, Benjamin (2009). "Integrated retrieval of research data and publication in digital libraries".

۱. دانشجوي كار شناسي ارشد دانشگاه پيام نور مشهد khosravi_hossin@yahoo.com

متغیرها، ایجاد معانی خاص داده ای که در حال حاضر با بازنمونه های معنایی تولید شده برای متون پژوهشی مطابقت کافی ندارند دارای پیچیدگی است اما پدیدار شدن انگاره مرتبط با علوم الکترونیکی (Gold, 2007) که به عنوان نشانه ترقی علم در نظر گرفته شده است توجه را به ایجاد زیر ساختهای جامع سخت افزاری و نرم افزاری و ایجاد شبکه های همکاری در جهت حمایت از فعالیتهای علمی پیشرفته که با گردآوری داده ها و یاداشتهای آزمایشگاهی آغاز و منجر به تولید سطح جدیدی از انتشارات علمی (مانند انتشارات الکترونیکی و مخازن دسترسی آزاد) گردید معطوف کرد و در همان زمان نتایج همه تحقیقات را برای محققان همکار قابل دسترس ساخت. بنابر این مدلها و روشهای علمی مورد نیاز است تا به طور یکدست ساختار و روابط معنایی انواع اطلاعات پژوهشی ارائه گردد و فرایندهای تطبیق و سازگاری به منظور شناسایی و پیوند اطلاعات مرتبط به هم جهت مستندسازی و هم برای بازیابی، تفسیر و استفاده مجدد از آنها تعریف گردد. همچنین این مدل ها مبنا و پایه ای هستند برای ویژگیهای پیشرفته ای مانند رایانش توزیعی، شبیه سازی و مصورسازی داده های ناهمگون و مجزا.

۲- پژوهشهای جاری

در حال حاضر تلاشهای بین المللی زیادی در جهت ایجاد دسترسی طولانی مدت به اطلاعات پژوهشی و استاندارد کردن قالب های آرشیوی صورت گرفته است . توصیفگرهای شیء دیجیتالی(DOI)[3] نمونه ای برای ایجاد توصیفگرهای دائمی هستند که مجموعه داده ها و انتشارات دیجیتالی را به مکانی که واقعا ذخیره شده اند، ارجاع می دهند . در سطح ساختاری، جامعه استانداردهائی را برای مستند کردن داده های اولیه طراحی می کند (مانند قالب DDI[4] از طرح داده های مدرک یا (SDMX)[5] که استانداردهایی که به عنوان استاندارد های فراداده ای برای اجاعات کتابشناختی فراهم شده اند (مانند مارک یا دوبلین کور[6] یا موجودیت های مرتبط برای مستند کردن نتایج فعالیت های پژوهشی (مانند CERIF[7] که مدلی برای سیستمهای اطلاعاتی پژوهشی است و طرح ها، موسسات، انتشارات، امکانات و تجهیزات پژوهشی، پروانه ثبت اختراعات و غیره را پوشش می دهد) . اما تاکنون این جوامع مختلف تنها به ارتباطات بی قاعده اکتفا کرده و استانداردها بیشتر بر تبادل فراداده ها تمرکز داشته اند (برای نمونه پروتکل های برداشت مانند (OAI_PMH)[8]. از قالب ابرداده ای که مدارک بر اساس آن ثبت می شود، جایابی های بین آنها و روش های رسمی برای جایابی طرحواره ای می توان به جای عناصری که کمترین تشابه و تطابق را در این قالب ها دارند استفاده نمود.

در سطح معنایی, برطرف کردن ناهمگونی خیلی پیچیده است به طوری که طرح های فراداده ای متفاوت از ابزارهای استاندارد جهت بازنمون محتوای معنائی استفاده نمی کنند، و همه معانی موجود در داده ها به صورت کامل بیان نمی شود. برای داده های اولیه مانند داده های مجموعه های زمانی و بررسی های پژوهشی انواع مختلفی از واژگان کنترل شده (مانند فهرست واژه ها و اصطلاحات و رده بندی ها) برای نمایه سازی محتوا استفاده می شود. در صورتی که اصطلاحنامه ها اغلب برای نمایه سازی داده های متنی (مانند انتشارات) استفاده می شوند. تطابق این واژگانهای متفاوت دشوار است و منجر به بروز اختلافاتی در ارائه مفاهیم معنائی

[3] DOI http://www.doi.org
[4] DDI http://www.ddialliance.org
[5] SDMX http://www.sdmx.org
[6] DUBLIN CORE http://www.dublincore.org
[7] CERIF http://www.eurocris.org/cerif/introduction
[8] OAI-PMH http://www.openarchives.org

و روابط استفاده شده جهت بیان انواع مختلف رابطه ها بین این مفاهیم در داخل هر لغت (برای مثال اصطلاحات عام، اصطلاحات خاص، مشابه و غیره) می شود. هر دو رویکرد نمایه سازی محتوا در زمینه های کاربردی متفاوتشان قابل توجیه هستند، اما چنانچه در یک سیستم بازیابی مورد استفاده قرار گیرند مشکلات تطابقی ایجاد می کنند. برای داده های اولیه، مرتبط ترین اطلاعات- نگرش علمی برای عبارت بندی یک سوال معین تنها به صورت تصادفی در خود سوال یا متغیر مربوطه رمزگذاری می شوند. (KRAUSE; STEMPFHUBER, 2005) این اطلاعات نمی تواند مستقیما در قالب شناسه های مناسب اصطلاح نامه که برای بازیابی متون مرتبط یا بالعکس مورد استفاده واقع می شوند قرار گیرند. در زمینه بازیابی متون، کاربران معمولا می توانند تا اندازه معینی از اشکالات و خشه ها را با پویش عنوان و چکیده نتایج رفع نمایند، اما در مورد بازیابی داده ها، در جائی که ربط در یک پژوهش برای بازاستفاده باید در سطحی از یک ترکیب شامل متغیر ها، روش نمونه گیری، اندازه نمونه، کدگذاری و غیره مورد قضاوت قرار گیرد لازم است تا نیازهای اطلاعاتی کاربران با دقت بیشتری شناخته و برآورده شود.

۳- مدلی برای یکپارچگی معنایی قالب های اطلاعاتی ناهمگون

مدل زیر یکپارچگی معنایی انواع ناهمگون اطلاعات موجود در کتابخانه های دیجیتالی را توصیف می نماید. این مدل به رفع ناهمگونی معنایی توجه کرده و قادر است که مشکلات ذکر شده فوق را حل کند. این مدل نه تنها محتویات معنایی انواع مختلف داده ها (مانند داده ها و انتشارات پژوهشی) را پوشش می دهد بلکه شامل معانی برای پیوند دادن داده ها با موجودیت های مرتبط با کل فرایند پژوهش می باشد. به طور خاص، این مدل شامل سه لایه می باشد و هر یک از لایه ها یک مشکل معنایی اختصاصی را حل می کند (نگاه کنید به تصویر ۱). در پاراگراف زیر سه لایه با جزئیات تشریح شده است .

تصویر شماره ۱ : مدل کامل

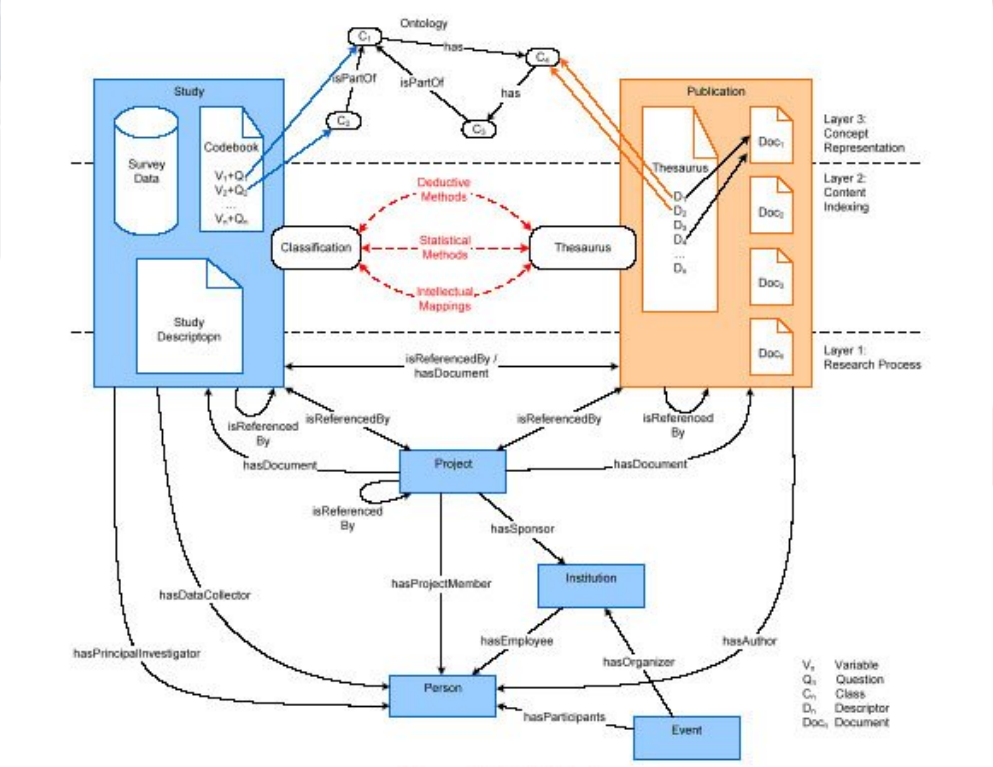

Figure 1: Full Model

۳-۱ لایه ۱ : فرایند پژوهش

این لایه (نگاه کنید به تصویر۲) کل فرایند پژوهش را منعکس می کند و روابط بین همه موجودیت ها (مانند اشخاص، سازمانها، برنامه های پژوهشی، طرح ها، نتایج، امکانات و توانمندیها، حق امتیازها) را نشان می دهد. بهر حال این لایه نشان دهنده چهار چوبی است که تحقیق در آن انجام و نتایج تحقیق حاصل شده است. این مدل بر اساس مدل هایی شبیه به استاندارد CERIF (قالب مشترک اطلاعات پژوهشی اروپا که توسط کمیسیون اروپا و euroCRIS توسعه داده شده یا هستی شناسی POLICY GRID بنا شده است. (CHORLEY & others, 2005) روابط موجود در این لایه، فرآیندهای استقرایی یا قیاسی را در حیطه پژوهش (برای مثال درباره نوشتن نتایج، برقراری پیوند نتایج به طرح ها، برقراری پیوند طرح های تکمیلی به برنامه های پژوهشی و غیره) فراهم می کند. از این روابط می توان برای کاوش اطلاعات مرتبط و طراحی هسته یک سیستم اطلاعاتی پژوهشی که به لایه های دیگر(در ادامه می بینید) متکی می باشند استفاده کرد.

معانی رمزگذاری شده در این لایه به کاهش ابهامات در فرایند بازیابی کمک می کند همانگونه که آنها اطلاعات زمینه ای برای نخستین درخواست های ارائه شده توسط کاربران در بافت کتابخانه های دیجیتالی ارائه می دهند (مانند متون یا داده های پژوهشی). آنها نه تنها توصیفگرهای دائمی و منحصر به فرد برای اشخاص در نقشهای مختلف (مانند نویسنده، محقق، مدیر پروژه و غیره) فراهم می کنند بلکه آنها اطلاعات تکمیلی (برای مثال در مورد اهداف برنامه های استراتژیک از برنامه های بودجه) ارائه می دهند که معمولا تنها در سطح موجودیت های اطلاعاتی منفرد بیان نمی شوند و می توانند برای پشتیبانی از راهبردهای کاوش کاربر نهائی مورد استفاده قرار گیرند.

تصویر شماره۲

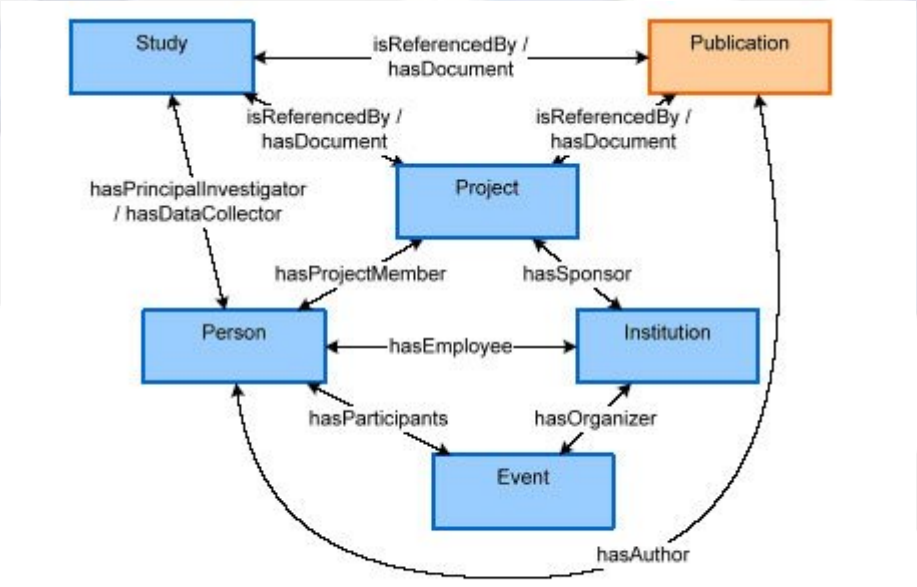

Figure 2: Layer for modelling the research process.

۳-۲ لایه۲: نمایه سازی محتوا

این لایه با معانی که از خود داده ها یا مدارک جایگزین (فراداده های همراه شامل نمایه سازی محتوا با کلید واژها، نشانه های رده بندی و غیره) برداشت می شوند ارتباط دارد. این لایه ناهمگونی بین واژگان نمایه

سازی به کاررفته در مجموعه های مختلف و در قالب های مختلف اطلاعات مانند رده بندی ها و اصطلاحنامه های تولید شده برای انتشارات را بررسی می کند. (نگاه کنید به تصویر ۳).

رویکردهای مقابله با ناهمگونی معنایی موجود بین واژه های نمایه سازی، شامل طرح های ذهنی (توافق دو جانبه)، روش های آماری و استنباطی، که معمولا جامعیت فرآیند را در بازیابی اطلاعات افزایش می دهد (Krause, 2004) و از کاربران با تغییر شکل خودکار درخواست ها برای یک نوع خاص اطلاعات (مانند انتشارات) به انواع دیگر اطلاعات (مانند داده های آماری) پشتیبانی می کند، بنابراین نیاز به یادگیری واژه های نمایه سازی جدید و فرمول بندی مجدد نیازهای اطلاعاتی در زمانهای مختلف را با استفاده از واژه های متفاوت برای یافتن اصطلاحات مناسب جستجو، رفع می کند. این انتقال خودکار می تواند به عنوان یک خدمت پیش زمینه ای شفاف و خودکار در طول فرایند پژوهش صورت بپذیرد. طرح هایی که واقعا در طول بازیابی استفاده واقع می شوند می توانند برای تامین اهداف اکتشافی و تکمیل و بسط مجموعه نتایج توسط کاربر مورد استفاده قرار گیرند .

تصویر شماره ۳

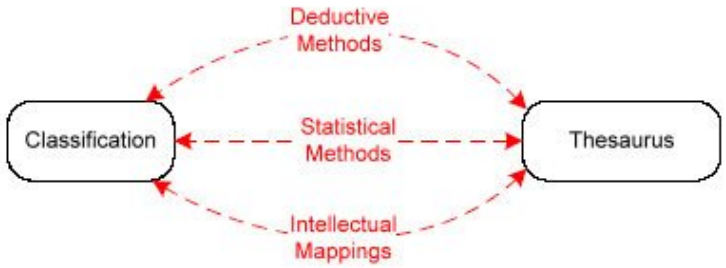

Figure 3: Layer for integrating the content indexing

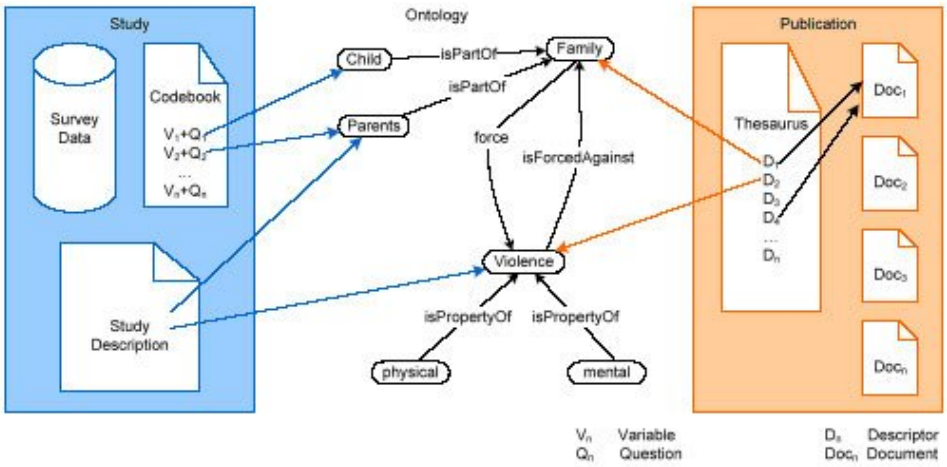

Figure 4: Layer for concept representation.

۳-۳. لایه۳: بازنمون مفهوم

لایه ۳ بالاترین لایه است(نگاه کنید به تصویر ۴) که تفاوتهای خاص بیان معنایی در بین اصطلاحنامه ها، رده بندی ها، دستورالعمل نامه ها را با تبدیل معانی پنهان موجود در مثلا داده های پژوهشی (یعنی نیت عالمانه جهت عبارت بندی یک سوال مشخص) به کلیدواژه های گویای ارزان تر مانند آنچه که در نمایه سازی انتشارات مورد استفاده قرار می گیرد دسته بندی می نماید. نمونه مشکلات ناشی از این شکاف در

بیان موقعیت هایی هستند که در آنها پژوهش های زیادی به دلیل جستجوهای کلید واژه های ساده در عبارات سوالی یا برچسبهای متغیر مرتبط شناخته شده اند، اما انجام تجزیه و تحلیل های عمیق از توصیفات پژوهشی نشان می دهد که آن پژوهش با نیاز اطلاعاتی کاربر تطابق ندارد و به درد وی نمی خورد. هستی شناسی در اینجا می تواند برای جنبه های معینی در عرصه علوم اجتماعی و هنر به عنوان پیوند بین معانی ساده تر از اصطلاحنامه برای پایگاهای متون (برای نمونه ارتباطات بین اصطلاحات خاص و عام) و جنبه های پیچیده موجود در سوالات پژوهشی و دستورالعمل نامه ها مورد استفاده قرار گیرد.

۴- نتیجه گیری و تحقیق بیشتر

مدل ارائه شده در اینجا سعی دارد که رویکردهای مکمل را در سازماندهی دانش و بازیابی اطلاعات در چارچوب کتابخانه های دیجیتالی و با وجود همه ناهمگونیهای موجود در سطوح معنایی و ساختاری آن ارائه کند. از این رو به دنبال غلبه بر کاستیهای رویکردهای فردی با یک دیدگاه یکپارچه است که تعداد زیادی از سیستمهای سازماندهی دانش سنتی (اصطلاحنامه ها) را قابل دسترس می سازد و به وسیله ترکیب آنها با هستی شناسی، مقداری کار مورد نیاز برای نمونه سازی تمام حوزه ها تا نمونه سازی – به عنوان اولین قدم – تنها بخش هایی از یک حوزه که اصطلاحنامه ها در آنها نمی توانند جهت حصول کیفیت مناسب بازیابی کارآیی داشته باشند کاهش می دهد.

محدوده کاربرد و بستر آزمون این مدل یکپارچه معنایی فهرست داده های GESIS[9] و یکپارچگی آن در دروازه علوم اجتماعی (sowiport.de)[10] که حاوی ۲/۵ میلیون رکورد از انتشارات، طرح ها، پرونده های سازمانی و غیره است، می باشد. در حالی که نتایج اولیه کارآیی لایه ۲ (نمایه سازی محتوا) نشان می دهد که جامعیت اطلاعات مرتبط قابل بهبود است (MAYR; PETRAS, 2008) اما ارتباطات معنائی شبیه اینها در لایه ۱و۲ در حال حاضر در مقیاس های مشابه ارزش گذاری نشده اند . این نکته مورد توجه ما در طرح های آینده خواهد بود.

منابع و مآخذ

CHORLEY,A; EDWARDS, P; HEILKEMA, F; PHILIP, L; and FARRINGTON, J. Developing Ontologies to Support eSocial Science: The PolicyGrid Experience. In Proceedings of the 4th International Conference on e-Social Science, Manchester,2008.

DELOS. The DELOS Network of Excellence on Digital Libraries: Recommendations and Observations for a European Digital Library (EDL). 4th DELOS Brain storming Wokshop on Digital Libraries, December 2005.

Gold, A.Cyberinfrastructure, Data and Libraries. Part 1&2. In D-Lib Magazine, 2007, Volume 13 Number 9/10.

[9] GESIS DATA CATALOGUE http://www.gesis.org/en/services/data/retrival-data-acess/data-catalogue
[10] SOWIPORT.DE http://www.sowiport.de

KARUSE, J. and STEMPFHUEBR, M. Nutzerseitige Integration sozialwissen-schatftlicher Text-und Datenintegration aus verteilten Quellen. In KONIG, C. et al. Datenfusion und Datenintegration: 6.Wissenschaftliche Tagung. Bonn, 2005, p. 141-158.

Krause, j. Standardization, Heterogenity and the Quality of Content Analyse: a key conflict of digital libraries and its solution. IFLA Journal: Official Journal of the International Federation of Library Associations and Institutions 30, 2004, No.4, S. 310-318.

MAYR, P. and PETRAS, V.Cross-concordances: terminology mapping and its effectiveness for information retrieval. IFLA World Library and Information Congress, 2008.

POLL, R. Nutzungsakalyse des systems der uberregionalen literature- und Informationsversorgung. Tile 1: Informationsverhalten und Informationsbedarf der Wissenschaft. In ZfBB 51, 2004, p. 59-75.