Data Mining and Deployment of Multilingual Iranian Cultural Thesaurus (ASFA) Dataset in the CRISP Framework

Saeedeh Akbari Daryan[1]

Abstract

Purpose: The Simple Knowledge Organization System (SKOS) is a widely used data model for sharing and linking knowledge organization systems on the web. It offers a cost-effective way to migrate existing knowledge organization systems to the Semantic Web. To integrate ASFA into the Semantic Web, the ASFA dataset needs to be converted and deployed as an RDF graph based on SKOS. To achieve this, the records in ASFA's Iran MARC format must be re-engineered. This study aims to re-engineer the ASFA dataset using data mining in the CRISP framework and deploy it on the open-source platform Skosmos.

Method: The study used the developmental-applied type of research and employed the CRISP-D.M. methodology, unsupervised type, and hierarchical clustering technique for data mining to start the project, we first needed to understand the business goal. This goal was to convert the ASFA dataset into the SKOS data model, creating an RDF graph. It was discovered that ASFA's heritage data comprises 1,880 records categorized into 18 fields, including education, literature, communication, economy, history, Sufism and mysticism, sociology, geography, law, psychology, linguistics, religion, political science, philosophy, technology, experimental science, librarianship and information, management, culture, and art. The data was prepared by identifying and correcting missing and outlier data and before starting the project, our team needed to fully comprehend the business's objective. The ultimate goal was to convert the ASFA dataset into the SKOS data model. This was done to better comprehend the business objective. Creating an RDF graph. The modeling stage utilized the hierarchical clustering technique macrocode in Excel to generate target feature values. The model was evaluated through a visual inspection technique and random sampling method. In the sixth step, Iran MARC data was converted to SKOS as an RDF graph using the SkosPlay tool, and the data was transferred to the Vocbench platform. ASFA Dataset was deployed on the Skosmos platform using the Turtle format.

Findings: The main finding of this study is the deployment and development of ASFA Dataset based on SKOS/RDF on the open source platform Skosmos at kosmos.nlai.ir. The total number of records increased to 11,880 records creating collection records for clustering. One of the important findings during the data preparation stage was the compilation of the mapping table between SKOS core elements and Iran MARC fields.

Conclusion: By integrating stages of methodologies used in the literature review within the CRISP framework, an innovative method was developed for converting thesauri into a lightweight ontology based on SKOS/RDF graph format.

Keywords

Data Mining, SKOS, Iran MARC, RDF Graph, Reengineering, Skosmos, ASFA Thesaurus

Citation: Akbari Daryan,S.(2023), Data Mining and Deployment of Multilingual Iranian Cultural Thesaurus (ASFA) Dataset in the CRISP Framework. *Librarianship and Information Organization Studies*,34(1): 58-82. Doi: 10.30484/NASTINFO.2023.3405.2209

DOI: 10.30484/NASTINFO.2023.3405.2209

Article Type: Research Article

Article history: Received: 19 Dec. 2022; Accepted: 7 Mar. 2023

1.Assistant Professor, National Library and Archives of I.R. of Iran, Tehran, Iran; sakbaridaryan@gmail.com

Publisher: National Library and Archives of I.R. of Iran

NASTINFO

داده‌کاوی و استقرار دادگان اصطلاحنامه چندزبانه فرهنگی ایران (اصفا) در چهارچوب کریسپ

سعیده اکبری داریان[1]

۱. استادیار سازمان اسناد و کتابخانه ملی ایران،
sakbaridaryan@gmail.com

چکیده

هدف: نظام ساده سازماندهی دانش (اسکاس) یک مدل داده‌ای رایج برای به‌اشتراک‌گذاری و پیونددهی نظام‌های سازماندهی دانش از طریق وب است. اسکاس، مسیر مهاجرت استاندارد و کم‌هزینه را برای انتقال نظام‌های سازمان دانش موجود به وب معنایی ارائه می‌دهد. پیوستن اصفا به جریان وب معنایی نیازمند تبدیل و استقرار دادگان اصفا براساس اسکاس در قالب گراف آر.دی.اف. است. به این منظور باید رکوردهای مبتنی بر مارک ایران مهندسی مجدد شوند. هدف پژوهش حاضر، مهندسی مجدد دادگان اصفا با داده‌کاوی آنها در چهارچوب کریسپ و استقرار آنها بر روی پلتفرم اسکاسموس است.

روش: این پژوهش از نوع توسعه‌ای - کاربردی است و از روش‌شناسی کریسپ-دی.ام.، از نوع بدون نظارت و خوشه‌بندی سلسله‌مراتبی برای داده‌کاوی استفاده شده است. در مرحله اول درک کسب و کار، هدف اصلی تبدیل دادگان اصفا به مدل داده‌ای اسکاس در قالب گراف آر.دی. اف. تعیین شد. در مرحله درک داده، داده‌های میراثی اصفا شامل ۱۱۰۰۶ رکورد ذخیره شده در قالب مارک ایران و شامل ۱۸ حوزه، آموزش و پرورش، ادبیات، ارتباطات، اقتصاد، تاریخ، تصوف و عرفان، جامعه‌شناسی، جغرافیا، حقوق، روان‌شناسی، زبان‌شناسی، دین، علوم سیاسی، فلسفه، فناوری و علوم تجربی، کتابداری و اطلاع رسانی، مدیریت و فرهنگ و هنر است. در مرحله سوم-آماده‌سازی داده- داده‌های مفقود و پرت شناسایی و ویرایش شد. برای انتخاب ویژگی‌ها در لایه پیش‌پردازش مهندسی داده، عناصر ضروری برای تبدیل به اسکاس شناسایی و جدول انطباق آنها با فیلدهای مارک ایران تدوین گردید. در مرحله مدل‌سازی، مقادیر ویژگی هدف با تکنیک خوشه‌بندی سلسله‌مراتبی و با استفاده از ماکروکد در اکسل تولید شد. ارزیابی مدل با تکنیک بررسی بصری و روش نمونه‌گیری تصادفی مورد تایید قرار گرفت. در مرحله ششم تبدیل داده‌های مارک ایران به اسکاس در قالب گراف آر.دی.اف. با استفاده از ابزار اسکاس‌پلی انجام و داده‌ها به بستر پلتفرم ووک‌بنچ انتقال یافت. با استفاده از قالب تورتل، دادگان اصفا در پلتفرم اسکاسموس مستقر شد.

یافته‌ها: یافته اصلی پژوهش، استقرار و توسعه دادگان اسکاس اصفا در پلتفرم منبع باز اسکاسموس به نشانی skosmos.nlai.ir است. مجموع رکوردها پس از ایجاد رکوردهای مربوط به حوزه و مجموعه برای خوشه‌بندی به ۱۱۸۸۰ رکورد افزایش یافت. در مرحله آماده‌سازی داده یکی از یافته‌های مهم، تدوین جدول انطباق بین عناصر هسته اسکاس و فیلدهای مارک ایران بود.

نتیجه‌گیری: در این پژوهش با بهره‌گیری از علم داده، روش نوآورانه‌ای برای داده‌کاوی دادگان اصطلاحنامه‌ای به‌کار رفت. روش‌شناسی‌های به‌کار رفته در ادبیات این پژوهش تنها در دو مرحله آماده‌سازی و استقرار و توسعه از شش مرحله به‌کار رفته در این پژوهش جا گرفتند.

کلیدواژه‌ها

داده‌کاوی، اسکاس، مارک ایران، گراف آر.دی.اف. مهندسی مجدد، اسکاسموس، اصطلاحنامه اصفا

استناد: اکبری داریان، سعیده(۱۴۰۲). داده‌کاوی و استقرار دادگان اصطلاحنامه چندزبانه فرهنگی ایران (اصفا) در چهارچوب کریسپ. *مطالعات کتابداری و سازماندهی اطلاعات*، ۳۴(۱): ۸۲-۵۸.

Doi:10.30484/NASTINFO.2023.3405.2209

فصلنامه مطالعات کتابداری و سازماندهی اطلاعات، ۳۴ (۱)، بهار ۱۴۰۲
DOI: 10.30484/NASTINFO.2023.3405.2209

نوع مقاله: پژوهشی
تاریخ دریافت: ۱۴۰۱/۰۹/۲۸
پذیرش: ۱۴۰۱/۱۲/۱۶

مقدمه

نظام سازماندهی دانش[1] به هر طرح رسمی برای مدیریت منابع اطلاعاتی اطلاق می‌شود. از جمله این طرح‌ها می‌توان به فایل‌های مستند[2]، سرعنوان‌های موضوعی[3]، تاکسونومی‌ها[4]، اصطلاحنامه‌ها[5]، شبکه‌های معنایی[6] و هستی‌نگاری‌ها[7] اشاره کرد (Theng, et al., 2009, 430) که با وجود ساختارها و عملکردهای متفاوت و روش‌های متنوع ارتباط با فناوری، وجه مشترک همه آنها این است که برای پشتیبانی از سازماندهی دانش و اطلاعات و با هدف تسهیل مدیریت و بازیابی اطلاعات طراحی شده‌اند (Mazzocchi, 2018). از دهه ۱۹۵۰ نظام‌های سازماندهی دانش در نخستین پایگاه‌های نمایه‌سازی و چکیده‌نویسی، خدمات اطلاع‌رسانی آنلاین، لوح‌های فشرده، فایل‌های ادوبی پی.دی.اف.، وب‌سایت‌های اچ.تی.ام.ال. و پایگاه‌های داده‌ای ایکس.ام.ال. به کار برده شده‌اند. اخیراً، این نظام‌ها سفر خود را برای پیوستن به جریان اصلی وب معنایی انجام داده و محصولات خود را به «داده‌های باز پیوندی» همراه با هستی‌نگاری‌های توسعه‌یافته در قرن بیست‌ویکم تبدیل کرده‌اند (Zeng & Mayr, 2019). اصطلاح هستی‌نگاری در سال‌های اخیر در علوم کامپیوتر و به‌ویژه در بازنمون دانش[8]، استدلال[9] و فناوری وب معنایی برجسته شده است (Davies, 2010).

یکی از اهداف وب معنایی که تیم برنرزلی[10] ابداع کرد، امکان جستجوی مفیدتر و یافتن آسان‌تر داده‌ها با استفاده از ابزارهای پژوهشی قوی‌تر و دقیق‌تر است. این امکان فراتر از جستجوی ساده بر اساس وجود کلمات در عنوان، چکیده یا متن کامل و از

1. Knowledge Organization System
2. Authority files
3. Subject headings
4. Taxnomies
5. Thesauri
6. Semantic network
7. Ontologies
8. Knowledge representation
9. Reasoning
10. Tim Berners-Lee

نظر معنایی انجام جستجو بر پایه محتوا می باشد. با این حال بلندپروازانه‌ترین هدف وب معنایی، گسترش وب از طریق یکپارچه‌سازی خودکار داده‌ها و اطلاعات بسیاری از منابع برخط، با پیاده‌سازی نرم‌افزاری با قابلیت استدلال خودکار است. در مفهوم اصلی وب معنایی عوامل نرم‌افزاری پردازش محتوا، یافتن اطلاعات از منابع مختلف و استدلال درباره داده‌ها و تولید خروجی را بر عهده دارند. براساس این مفهوم‌سازی، هستی‌نگاری‌ها مناسب‌ترین ابزار برای فراتررفتن از مرزهای راهبردهای سنتی برای یافتن و دسترسی به اطلاعات درنظر گرفته شده‌اند (Biagetti, Maria Teresa, 2021). در اصل وب معنایی یک وب مجزا نیست، بلکه توسعه‌ای از وب فعلی است که در آن کامپیوترها داده‌ها را به‌جای انسان تفسیر می‌کنند. از نظر امیرحسینی (۱۴۰۱الف) هدف غایی وب معنایی، کاربرد آن در راستای هوشمندسازی وب در تحقق تعامل بین انسان و ماشین از طریق شکل‌گیری روابط معنایی بین مفاهیم در حوزه‌های گوناگون موضوعی در بستر هستی‌نگاری‌ها تفسیر می‌شود. او معتقد است این دیدگاه در سازماندهی اطلاعات و به ویژه دانش، نقش اساسی دارد و عامل ذخیره مناسب و بازیابی مؤثر اطلاعات و دانش به منظور شکل‌گیری ارتباط تعاملی پیش‌گفته می‌شود.

با افزایش اهمیت و جایگاه وب معنایی مبحث بقای اصطلاحنامه‌ها در وب معنایی مطرح و الگوهایی نیز برای انتقال اصطلاحنامه‌های مبتنی بر اصطلاح[1] به مدل‌های داده‌های مبتنی بر مفهوم[2] و در نهایت به هستی‌نگاری‌های سبک[3] ارائه شد (Villazón-Terrazas, Suárez-Figueroa, & Gómez-Pérez, 2010). همان گونه که دیویس[4] اشاره می‌کند مفهوم هستی‌نگاری سبک در حوزه کامپیوتر قابل درک است که در آن هستی‌نگاری مصنوع نرم‌افزاری و یک مدل قابل پردازش کامپیوتری وجود دارد. هستی‌نگاری‌های سبک در این زمینه، ساده‌ترین رسمی‌سازیِ ساده‌ترین مدل را فراهم می‌کنند. منطق این است که هستی‌نگاری‌های ساده اغلب مناسب‌تر و مقرون به‌صرفه‌ترند. درک، تطبیق، مدیریت، به‌روزرسانی و استفاده از آنها نیز آسان‌تر است. هستی‌نگاری‌های سبک می‌توانند در محیط‌های پردازشی که مقیاس و عملکرد آنها حیاتی است مانند پایگاه داده‌ها و موتورهای جستجوی بسیار بزرگ زنده بمانند. در سال‌های اخیر، تعدادی از پژوهشگران، هستی نگاریی‌های سبکی را برای استفاده در تعدادی از سناریوهای کاربردی مختلف ساخته‌اند. آنها اغلب سازوکارهایی را برای نگاشت بین صورت گرایی‌های دیگر در طیف معنایی به یک هستی‌شناسی سبک ارائه کرده‌اند. این امر از آنجایی مهم است که اولاً، مزایای دقت، معناشناسی رسمی و بهبود پردازش‌پذیری ماشین را به طرح‌های قبلی اضافه می‌کند و ثانیاً، راهی

1. Term-based thesaurus
2. Concept-based thesaurus
3. Lightweight ontology
4. Davis

برای تولید هستی‌شناسی‌های دامنه با هزینه معقول بر اساس کارهای از قبل موجود معمولاً با حداقل درجه‌ای از اجماع جامعه ارائه می‌کند(Davies, 2010).

ون آسم[1] و همکاران (۲۰۰۶)، روشی را برای تبدیل اصطلاحنامه به یک هستی‌نگاری سبک نظام ساده سازماندهی دانش (ازاین‌پس اسکاس) ارائه می‌کنند. اساسا، اسکاس به عنوان یک فرامدل برای نمایش اصطلاحنامه‌ها در آر.دی.اف.[2] به‌کار می‌رود. هم‌اکنون اغلب اصطلاحنامه‌های برجسته نسخه اسکاس اصطلاحنامه خود را ارائه می‌دهند و ابزارهای اصطلاحنامه در حال تحول‌اند تا از این ساختار پشتیبانی کنند. داده‌های پیوندی برای اصطلاحنامه‌ها خوب بوده است. زیرا می‌توان اصطلاحنامه‌ها را با سایر نظام‌های سازماندهی دانش که دانش مشابهی را به اشتراک می‌گذارند، پیوند داد. بنابراین داده‌های پیوندی، سودمندی ابزارهای اشتراک‌گذاری دانش را افزایش می‌دهند (Martínez-González & Alvite-Diez, 2019). اسکاس یک مدل داده‌ای رایج برای به‌اشتراک‌گذاری و پیونددهی نظام‌های سازماندهی دانش از طریق وب است. بسیاری از نظام‌های سازماندهی دانش، مانند اصطلاحنامه‌ها دارای ساختار مشابهی هستند و در برنامه‌های مشابهی به‌کار گرفته می‌شوند. اسکاس با گرفتن این شباهت‌ها امکان به‌اشتراک‌گذاری داده‌ها و فناوری را در برنامه‌های مختلف فراهم می‌کند. مدل داده‌ای اسکاس، مسیر مهاجرت استاندارد و کم‌هزینه‌ای را برای انتقال نظام‌های سازمان دانش موجود به وب معنایی ارائه می‌دهد. همچنین اسکاس از زبانی سبک و قابل درک برای توسعه و به‌اشتراک‌گذاری نظام‌های سازماندهی دانش برخوردار است. اسکاس ممکن است به تنهایی یا در ترکیب با زبان‌های بازنمون رسمی دانش مانند زبان هستی‌نگاری وب استفاده شود (Miles, A., & Bechhofer, S. 2009). در نهایت، انطباق و اتصال دادگان یک اصطلاحنامه به اصطلاحنامه‌های دیگر در قالب اسکاس/آر.دی.اف. میانکنش‌پذیری و ایجاد داده‌های پیوندی را به دنبال دارد. در این صورت، هنگام جستجو، مدارک متصل به داده‌های چندین اصطلاحنامه بازیابی می‌شود و در صورت چندزبانه بودن اصطلاحات، مدارک به زبان‌های مختلف بازیابی خواهد شد.

سازمان اسناد و کتابخانه ملی ایران[3] سال‌هاست نظام‌های سازماندهی دانش مانند سرعنوان‌های موضوعی، گسترش‌های رده‌بندی دیویی و کنگره و نیز اصطلاحنامه‌ها را براساس استانداردهای بین‌المللی تدوین و پشتیبانی می‌کند. *سرعنوان‌های موضوعی فارسی*، *سرعنوان‌های موضوعی کودکان*، *اصطلاحنامه چندزبانه فرهنگی ایران* (اصفا) و *اصطلاحنامه پزشکی* همگی در اپک کتابخانه ملی ایران[4] بدون امکان تفکیک

1. Van Assem
2. RDF
۳. به منظور اختصار از این پس، کتابخانه ملی ایران نوشته خواهد شد.
4. Opac.nlai.ir

نظام‌های مختلف هنگام جستجو به شکلی یکسان نمایش داده می‌شوند و امکان نمایش سلسله‌مراتبی آنها وجود ندارد. بنابراین برای حیات نظام‌های سازمان‌های دانش کتابخانه ملی ایران در جریان وب معنایی، مهندسی مجدد داده‌های آن برای بهره‌گیری از استانداردهای این حوزه امری ضروری است. اصفا با شمول زبان‌های فارسی، انگلیسی، عربی، فرانسه، روسی و خط‌های اصطلاحات تاجیکی به خط فارسی و اصطلاحات تاجیکی به خط سیریلیک، گزینه مناسبی است تا با قرار گرفتن آن در جریان وب معنایی میانکنش‌پذیری، استفاده مجدد و امکان پیوند آن با اصطلاحنامه‌های مرتبط به زبان‌های مختلف امکان پذیر شود.

واژه‌نامه مریام وبستر[1] مهندسی مجدد را طراحی مجدد یا سازماندهی مجدد عملیات یک سازمان برای بهبود کارایی تعریف می‌کند. واژه‌نامه مک‌گروهیل[2] (۲۰۰۳) **نیز** مهندسی مجدد را کاربرد فناوری و علم مدیریت برای اصلاح سیستم‌ها، سازمان‌ها، فرایندها و محصولات موجود به منظور مؤثرتر، کارآمدتر و پاسخگو کردن آنها می‌داند. واژه‌نامه مک‌گروهیل (۲۰۰۲) خاطرنشان می‌کند اصطلاح «مهندسی مجدد» به معنای نوعی بازسازی یا بهسازی یک محصول از قبل مهندسی شده است و مفهوم تعمیر و نگهداری یا نوسازی از آن تعبیر می‌شود. مهندسی مجدد را می‌توان مهندسی معکوس نیز تعبیر کرد، که در آن ویژگی‌های یک محصول مهندسی‌شده از قبل شناسایی می‌شود؛ به طوری که ممکن است محصول را اصلاح یا از آن استفاده مجدد کرد. در ذات این مفاهیم دو جنبه اصلی مهندسی مجدد وجود دارد: محصول یا سیستم ارائه‌شده به کاربر را برای افزایش قابلیت اطمینان[3] یا نگه‌داشت‌پذیری[4] یا برای برآوردن نیازهای جدید کاربران سیستم بهبود می‌بخشد و این، درک خود سیستم یا محصول را افزایش می‌دهد. این تفسیر از مهندسی مجدد کاملاً بر محصول متمرکز است.

امیرحسینی (۱۴۰۱) مهندسی مجدد اصطلاحنامه‌ها را بازطراحی و بازاندیشی فرایندها در اصلاح و توسعه نظام‌های سازماندهی دانش سنتی می‌داند. او همچنین هدف آن را یافتن بهترین فرایند با تکیه بر معیارهایی مانند هزینه، کیفیت خدمات و سرعت در راستای بهبود کارایی، افزایش بهره‌وری و افزایش عملکرد در سازماندهی دانش می‌داند و معتقد است مهندسی مجدد اصطلاحنامه‌ها از طریق تجزیه و تحلیل مسائل موجود، ترسیم اهداف و برنامه مورد نظر و ارائه راهکارهای مناسب قابل انجام است.

براساس آنچه گذشت، دادگان اصفا در کتابخانه ملی ایران با استاندارد مارک ایران ذخیره شده است و به دلیل قدیمی بودن معماری اپک کتابخانه ملی ایران[5] علاوه

1. Merriam-Webster
2. McGraw-Hill
3. Reliability
4. Maintainability
۵. معماری اپک کتابخانه ملی ایران به سال ۱۳۸۵ برمی‌گردد و از آن زمان تاکنون توسعه نیافته است.

بر نبود نمایش درختی اصطلاحات، در قالب یک پایگاه داده مستقل، قابل جستجو نیست. اصفا به دلیل دارا بودن اصطلاحات به چندزبان و چندخط، گزینه بسیار مهمی برای آغاز پیوستن کتابخانه ملی ایران به جریان وب معنایی است. این امر، نیازمند تبدیل و انتشار دادگان اصفا براساس اسکاس در قالب گراف آر.دی.اف. است. هدف پژوهش حاضر، مهندسی مجدد دادگان اصفا با داده‌کاوی آنها در چهارچوب کریسپ و استقرارشان بر روی پلتفرم اسکاسموس است.

پیشینهٔ پژوهش

مرور ادبیات بین المللی نشان می‌دهد، برای تبدیل اصطلاحنامه به اسکاس/ آر.دی.اف. روش‌های متعددی به کار رفته است:

مایلز، راجرز و بکت (۲۰۰٤) روش سه مرحله‌ای را برای تبدیل اصطلاحنامه به نسخه اولیه اسکاس پیشنهاد داده‌اند: ۱) کدگذاری آر.دی.اف.۲) بررسی خطا و اعتبارسنجی و ۳) انتشار آر.دی.اف. در وب. این روش مبتنی بر دو الزام است: الف) تبدیل یک اصطلاحنامه به مدل اسکاس با هدف پشتیبانی از میان‌کنش‌پذیری اصطلاحنامه؛ ب)حفظ تمام اطلاعات کدگذاری‌شده در اصطلاحنامه. در مرحله اول تبدیل، اصطلاحنامه‌ها از منظر «ساختار غیر استاندارد» یا «ساختار استاندارد» تفکیک می‌شوند. اصطلاحنامه با ساختار استاندارد بر اساس استاندارد ایزو ۲۷۸۸ است. چنین اصطلاحنامه‌هایی را می‌توان بدون از دست دادن اطلاعات به نمونه‌هایی از طرح‌واره اسکاس تبدیل کرد. اصطلاحنامه‌هایی با ساختار غیراستاندارد دارای ویژگی‌های ساختاری‌اند که در استاندارد ایزو ۲۷۸۸ توصیف نشده‌اند. توصیه می‌شود در این موارد توسعه اسکاس با به‌کارگیری rdfs:subClassOf و rdfs:subPropertyOf برای پشتیبانی از ویژگی‌های غیر استاندارد انجام شود. زیرا این راه‌حل تضمین می‌کند که هر دو الزامات روش برآورده شده‌اند. مرحله دوم شامل بررسی خطا و اعتبارسنجی با استفاده از اعتبارسنج‌های کنسرسیوم وب است، در این پژوهش درباره مرحله سوم خیلی بحث نشده است. آنها این روش را برای سه اصطلاحنامه «خدمات اطلاع‌رسانی روابط عمومی‌استرالیایی[1]»، اصطلاحنامه میراثی انواع هواپیماها[2] و اصطلاحنامه عمومی‌چندزبانه محیطی[3] اجرا کردند. اصطلاحنامه‌های اول و دوم دارای ساختار استاندارد بودند و با روش بالا کدگذاری آر.دی.اف. شدند اما به دلیل اینکه اصطلاحنامه سوم ساختار استاندارد نداشت، روش فوق قابل اجرا نبود.

ون آسم[4] و دیگران (۲۰۰٦) روشی را برای تبدیل نیمه خودکار اصطلاحنامه به

1. The Australian Public Affairs Information Service Thesaurus (APAIS)
2. Aircraft Type Thesaurus (ATT)
3. GEneral Multilingual Environmental Thesaurus (GMET)
4. Assem

هستی‌نگاری سبک در چهار مرحله شرح می‌دهند: ۱) آماده‌سازی ۲) تبدیل نحوی ۳) تبدیل معنایی و ٤) پذیرش استاندارد بین‌المللی اسکاس. در مرحله اول، تحلیلی از اصطلاحنامه و قالب دیجیتالی آن انجام می‌شود که از این در مرحله دوم برای تبدیل نحوی به آر.دی.اف. استفاده می‌شود، پس از آن به مدل‌سازی رایج تر مورد استفاده در آر.دی.اف/اُ.دبلیو.ال. در مرحله سوم تبدیل می‌شود. در آخرین مرحله، فرامدل آر.دی.اف/اُ.دبلیو.ال. آن اصطلاحنامه با اسکاس مطابقت[1] داده می‌شود. این روش مبتنی بر دو الزام است: ۱) حفظ معنایی اصطلاحنامه و ۲) پالایش گام به گام فرامدل آر.دی.اف/اُ. دبلیو.ال. اصطلاحنامه.

حاصل پژوهش باربوسا[2] و دیگران (۲۰۲۱) پیشنهاد روش شش مرحله‌ای است:۱) انتخاب و استخراج داده ۲) انطباق و تبدیل داده‌ها ۳) ارائه آر.دی.اف. ٤) سریال‌سازی[3] آر.دی.اف./ تورتل[4] ۵) مجوز داده‌های باز و ٦) پرس‌وجوی انتشار. نخستین مرحله، انتخاب واژگان کنترل‌شده از رکوردهای مستندی است که معمولاً در پایگاه داده‌های مستند ذخیره می‌شوند. سپس لازم است قالب اصطلاحنامه‌ها تجزیه و تحلیل شود تا ساختار مفاهیم مشخص گردد. استخراج رکوردهای مستند باید در قالب ایکس.ام.ال. صورت پذیرند تا در مرحله مطابقت داده‌ها و تبدیل استفاده شوند. در مرحله دوم، از فایل‌های ایکس.ام.ال. استخراج شده به‌عنوان ورودی برای انطباق و فرایند تبدیل رکوردهای مستند به مدل داده‌ای اسکاس با سازوکار تبدیل زبان صفحه سبک توسعه‌پذیر (ایکس.اس.ال.تی.)[5] استفاده می‌شود. در مرحله سوم به منظور تبدیل داده‌ها به اسکاس/ آر.دی.اف. و ویرایش سه‌گانه آر.دی.اف.، می‌توان از ابزارهای مدیریت اسکاس بهره برد؛ مانند اسکاس ای.دی.[6]، اسکاس تو اُ.دبلیو.ال.[7] در ادامه، فایل‌های آر.دی.اف. باید براساس انواع نحوهای آر.دی.اف. به صورت یک سری نویسه کدگذاری شوند. سریال‌سازی به سه‌گانه‌ها این امکان را می‌دهد تا به آسانی توسط سایر برنامه‌های کاربردی وب مورد استفاده مجدد قرار گیرند که معمولا دستورالعمل‌های استفاده از قالب آر.دی.اف. را پیشنهاد می‌کنند. مرحله پنجم، بهره گیری از انواع مجوزهای استفاده می‌باشد که پیشنهاد آنها برای انتشار داده‌های باز پیوندی در حوزه عمومی‌سی.سی.زیرو[8] است. در مرحله آخر، پس از نمایش در مدل داده‌ای اسکاس و ذخیره در دادگان‌های آر.دی.اف، واژگان کنترل‌شده را می‌توان برای استفاده در سایر برنامه‌های کاربردی وب معنایی در وب در دسترس قرار داد.

اکبری داریان و انتهایی (۱۳۹۹) نیز برای تبدیل داده‌های اصفا مبتنی بر مارک ایران به اسکاس/ آر.دی.اف، مدلی به شرح زیر ارائه کردند:

1. Map
2. Barbosa
3. Serialisation
4. Turtle
5. eXtensible Stylesheet Language for Transformation (XSLT)
6. SKOS ED.
7. Skos2OWL
8. CC0

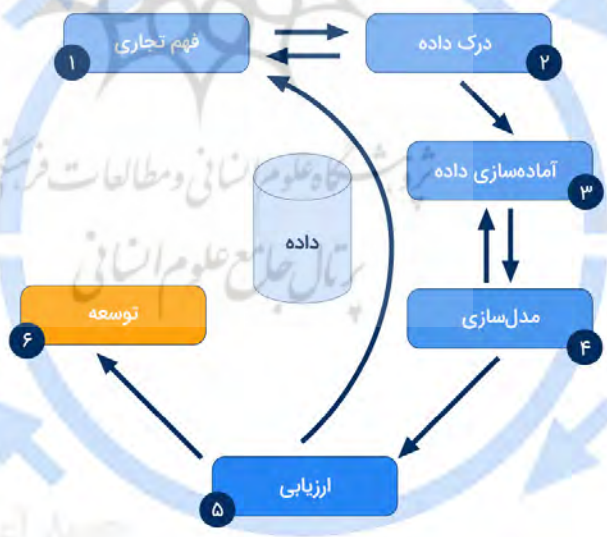
۱) اصلاح و ویرایش‌های محتوایی شناسایی‌شده در آزمون داده‌های اصفا ۲) اضافه کردن مشخصه‌ای برای مفاهیم رأس در اصفا در محیط مارک ایران ۳) تبدیل خروجی ایزو به اکسل سازگار با اسکاس پلی[1] ٤) آزمون داده‌ها در ووک بنچ[2]، ۵) تجزیه یو.آر.آی.[3] در اسکاسموس[4].

این بررسی‌ها نشان می‌دهد، ادبیات بین‌المللی از سال ۲۰۰٦ مطالعاتی را در مورد روش‌های مختلف برای تبدیل اصطلاحنامه‌ها به اسکاس/آر.دی.اف.، معمولا با استفاده از روش‌های نیمه خودکار ارائه می‌دهد که این مدل‌ها غالبا تبدیل نحوی است. تکامل روش‌ها در گذر زمان ملموس است اما روی نحوه آماده‌سازی داده‌ها -که بخش مهمی از روش تبدیل به اسکاس است- توضیح داده نشده‌است؛ این درحالی است که پژوهش داخل کشور درباره اصفا نشان می‌دهد اشکالات موجود در داده‌ها مانع از استقرار آنها در پلتفرم اسکاسموس شده است. بنابراین وزن گام مربوط به آماده‌سازی داده در این پروژه سنگین‌تر از سایر گام‌ها خواهد بود. به نظر می‌رسد داده‌کاوی[5] می‌تواند به شناسایی دقیق خلاءها و مشکلات دادگان اصفا منجر شود. اصولا داده‌کاوی مرحله‌ای در فرایند کشف دانش در پایگاه‌های داده[6] است که از طریق آن دارایی‌های داده‌ای پردازش و تجزیه و تحلیل می‌شوند تا بینش‌هایی برای کمک به تصمیم‌گیری به دست آید. فرایند کشف دانش از داده‌های نگهداری شده در سیستم‌های مدیریت داده سازمان یا انبارهای داده[7] سرچشمه می‌گیرد. مراحل کشف دانش شامل انتخاب داده، پردازش، تبدیل، داده‌کاوی، تفسیر و ارزیابی است که منجر به کشف اطلاعات یا دانش جدید می‌شود. با این اوصاف برای شناسایی و تجزیه و تحلیل داده‌ها می‌توان از روش‌های داده‌کاوی بهره گرفت و دانش پنهان موجود در داده‌ها را استخراج کرد. روش‌های داده‌کاوی به حوزه کاربرد آنها بستگی دارد. الزامات مهم این است که داده‌های جمع‌آوری شده باید مرتبط و باکیفیت باشند (Haravu, & Neelameghan, 2003). تا به امروز، بسیاری از روش‌ها و مدل‌های فرآیند داده‌کاوی و کشف دانش - با درجات مختلف موفقیت - توسعه یافته‌اند. ماریسکال و همکاران (۲۰۱۰) بررسی جامعی بر روی فرآیند داده‌کاوی و کشف دانش انجام دادند. آنها نقشه تکاملی از ١٤ مدل ارائه کردند که در میان آنها کریسپ- دی.ام. دارای رویکرد مرکزی است. کریسپ- دی.ام با سهم ٤٣ درصدی در آخرین نظرسنجی کی.دی. ناگتز[8] (جامعه شناخته شده در فرآیند داده‌کاوی و کشف دانش) محبوب‌ترین روش برای پروژه‌های تجزیه و تحلیل، داده‌کاوی و علم داده است (Piatetsky-Shapiro, 2014).

1. SKOS Play
2. VocbBench
3. Resolve URI
4. Skosmos .
5. Data mining
6. knowledge discovery in the databases (KDD)
7. Data warehouse
8. KDnuggets

روش پژوهش

پژوهش حاضر از نوع توسعه‌ای – کاربردی است. از روش‌شناسی کریسپ–دی.ام.[1]، از نوع بدون نظارت[2] و خوشه‌بندی سلسله‌مراتبی[3] برای داده‌کاوی استفاده شده است. معمولا در پروژه‌های علم داده از روش‌شناسی فرایند استاندارد صنعت متقابل برای داده‌کاوی (کریسپ- دی.ام.) استفاده می‌شود. این یک مدل فرآیندی با شش مرحله درک کسب و کار، درک داده، آماده‌سازی داده، مدل‌سازی، ارزیابی و استقرار است که به طور طبیعی، چرخه حیات علم داده را توصیف می‌کند (شکل ۱). هدف خوشه‌بندی (تحلیل خوشه‌ای)، سازمان‌دهی مجموعه‌ای از اقلام داده در خوشه‌هاست؛ به‌گونه‌ای که اقلام درون یک خوشه بیشتر «شبیه» به یکدیگر باشند تا اقلام موجود در دیگر خوشه‌ها. این مفهوم شباهت را می‌توان به روش‌های بسیار متفاوتی- با توجه به هدف مطالعه- مفروضات خاص حوزه و دانش قبلی از مسئله بیان کرد. خوشه‌بندی معمولاً زمانی انجام می‌شود که هیچ اطلاعاتی در مورد عضویت اقلام داده‌ای در دسترس نباشد. به همین دلیل، خوشه‌بندی به طور سنتی به عنوان بخشی از یادگیری بدون نظارت است. در پژوهش حاضر، ابتدا اشاره مختصری به مراحل روش‌شناسی کریسپ و سپس به تفضیل این مراحل پرداخته می‌شود.



شکل ۱- روش‌شناسی کریسپ

1. Cross Industry Standard Process for Data Mining (CRIPSP-DM)
2. Unsupervised
3. Clustering

در نخستین مرحله روش کریسپ، مسئله و اهداف پژوهش تعریف می‌شود (درک کسب وکار). مرحله بعد، جمع‌آوری و تجزیه و تحلیل داده‌هاست (درک داده). در مرحله سوم، داده‌ها به منظور تجزیه و تحلیل، تمیز و پیش‌پردازش می‌شود (آماده‌سازی داده)؛ این امر می‌تواند شامل حذف موارد تکراری یا داده‌های نامربوط و تبدیل ویژگی‌ها[1] به قالب مناسب برای تجزیه و تحلیل باشد. در مرحله چهارم از میان تکنیک‌های مختلف داده‌کاوی مدل در داده‌ها اعمال می‌شود (مدل‌سازی). مرحله پنجم نحوه ارزیابی عملکرد مدل در دستیابی به اهداف پروژه است (ارزیابی). در مرحله نهایی مدل توسعه داده و داده‌ها در یک پلتفرم مستقر می‌شود(توسعه).

مرحله اول: درک کسب و کار

هدف اصلی پژوهش، تبدیل دادگان اصفا به مدل داده‌ای اسکاس در قالب گراف آر.دی. است. تاریخچه *اصطلاحنامه چندزبانه فرهنگی ایران* به سال ۱۳۷۲ برمی‌گردد. نخستین ویراست آن با عنوان «*اصطلاحنامه فرهنگی فارسی (اصفا)*» در سال ۱۳۷۶ توسط سازمان مدارک فرهنگی انقلاب اسلامی، سپس در سال ۱۳۸۵با عنوان «*اصطلاحنامه فرهنگی سه زبانه فارسی- انگلیسی – عربی*» در کتابخانه ملی ایران منتشر شد. در همان سال تمام اصطلاحات اصفا از نرم‌افزار اولیه‌ای که توسط شرکت پارس‌آذرخش برای اصطلاحنامه‌ها طراحی شده بود، به نرم‌افزار جامع کتابخانه ملی ایران (رسا) در قالب مارک ایران انتقال یافت و همزمان اصطلاحات جدید در دو نرم‌افزار وارد می‌شد. زیرا خروجی رسا قابلیت انتشار اصطلاحات به شکل اصطلاحنامه چاپی و با ساختارهای رده‌ای و درختی را ندارد. ویراست دوم اصطلاحنامه سه زبانه در سال ۱۳۹۲ منتشر و این آخرین نسخه چاپی این اصطلاحنامه بود. پس از آن اصطلاحات جدید فقط در نرم‌افزار رسا وارد شد. به منظور افزایش کاربران بالقوه این اصطلاحنامه در کشورهای هدف، در سال ۱۳۹۶ برابر نهاده‌های روسی و تاجیکی نیز به آن اضافه شد. همچنین افزودن اصطلاحات مرجح فارسی به خط سیریلیک و اصطلاحات تاجیکی به خط سیریلیک و نیز به خط فارسی بر غنای این اصطلاحنامه افزود. از آنجایی که میلیون‌ها سند و منابع غیرکتابی در کتابخانه ملی ایران بر اساس اصفا نمایه‌سازی می‌شود؛ به‌صورت مداوم اصطلاحات جدید از سوی متخصصان نمایه‌سازی پیشنهاد و بدین ترتیب این اصطلاحنامه روزآمد می‌گردد. داده‌های اصفا در رسا دودسته‌اند؛ بخشی از آنها حاصل انتقال از سیستم قبلی و بخش دیگر اطلاعات جدید است. بررسی اولیه نیز نشان می‌دهد اطلاعات پرت (نویز یا خارج از محدوده[2]) و داده‌های مفقود[3] در دادگان دسته اول قابل ملاحظه است.

1. Features
2. Outliers data
3. Missing data

مرحله دوم: درک داده

داده‌ها، مشاهدات یا اندازه‌گیری‌ها (پردازش نشده یا پردازش شده) هستند که به صورت متن، اعداد یا چندرسانه‌ای نمایش داده می‌شوند. دادگان، مجموعه‌ای ساختاریافته از داده‌هاست که عموماً با یک مجموعه منحصر به فرد مرتبط است. منظور از داده در این پژوهش ۱۱۰۰٦ رکورد اصفا می‌باشد که در قالب مارک ایران در نرم‌افزار رسا ذخیره شده است. دادگان اصفا شامل ۱۸ حوزه، آموزش و پرورش، ادبیات، ارتباطات، اقتصاد، تاریخ، تصوف و عرفان، جامعه‌شناسی، جغرافیا، حقوق، روان‌شناسی، زبان‌شناسی، دین، علوم سیاسی، فلسفه، فناوری و علوم تجربی، کتابداری و اطلاع‌رسانی، مدیریت، و فرهنگ و هنر است. هر حوزه یک دادگان مستقل محسوب می‌شود.

جدول۱- نمایش یک رکورد اصفا قبل از اجرای پروژه

ردیف	داده‌ها	شماره فیلد (تگ)	برچسب فیلد
۱	01739cx j2200445 450	0	برچسب رکورد
۲	153062	1	شناسگر رکورد
۳	##$a20060906apery50 fa1	100	داده‌های کلی پردازش
٤	##$aper	101	زبان موجودیت
٥	##$basfa$9تا_اص	152	قواعد
٦	##$a٦$تاریخی آثارa00$7fa$8per	250	شناسه دستیابی مستند- موضوع
۷	##$aHistorical Remains$6a01$7ba$8eng	250	شناسه دستیابی مستند- موضوع
۸	##$a<٦$$التاريخية=آثار>الآثارa01$7fa	250	شناسه دستیابی مستند- موضوع
۹	##$апамятники$6a01$8rus	250	شناسه دستیابی مستند- موضوع
۱۰	##$аёдгориҳои таърихӣ$6a01$7ca$8tgk	250	شناسه دستیابی مستند- موضوع
۱۱	##$a٦$یادگاری‌های تأریخیa01$7fa$8tgk	250	شناسه دستیابی مستند- موضوع

ردیف	داده‌ها	شماره فیلد (تگ)	برچسب فیلد
۱۲	##$aocopэ торЙхЙ$6a01$7ca$8per	250	شناسه دستیابی مستند- موضوع
۱۳	##$aArchaeological Objects$7ba$6a02	450	شناسه دستیابی دیگر- موضوع
۱٤	##$aArchaeological Remains$7ba$6a00	450	شناسه دستیابی دیگر- موضوع
۱٥	##$aHistorical Objects$7ba$6a03	450	شناسه دستیابی دیگر- موضوع
۱٦	##$a<۷$القديمة>آثار=الآثار|fa$6a00	450	شناسه دستیابی دیگر- موضوع
۱۷	##$a۷$آثار باستانیfa	450	شناسه دستیابی دیگر- موضوع
۱۸	##$a<۷$الاثرية>اشياء=الاشياء|fa$6a02$8ara	450	شناسه دستیابی دیگر- موضوع
۱۹	##$a<۷$التاريخية>اشياء=الاشياء|fa$6a03$8heb	450	شناسه دستیابی دیگر- موضوع
۲۰	##$a۷$اشیای باستانی|fa$8per	450	شناسه دستیابی دیگر- موضوع
۲۱	##$a۷$اشیای تاریخی|fa$8per	450	شناسه دستیابی دیگر- موضوع
۲۲	##$a٥$باستان‌شناسیh	550	شناسه دستیابی مرتبط- موضوع
۲۳	##$a٥$۳۳۱٥۳۳۰٦۳آرامگاه‌هاh6a00$7fa$8per	550	شناسه دستیابی مرتبط- موضوع
۲٤	##$aبرج‌های تاریخی3186050$5h6a00$7fa$8per	550	شناسه دستیابی مرتبط- موضوع
۲٥	##$a٥$۳۳۰۹۳٤٥سنگ قبرهاh7fa$8per	550	شناسه دستیابی مرتبط- موضوع
۲٦	##$a٥$۳۲۸۰۲۷۲قلعه‌هاh7fa$8per	550	شناسه دستیابی مرتبط- موضوع
۲۷	##aIRc20211017113926.0130145.0130103.0$bکتابخانه ملی	801	مبدأ اصلی

ردیف	داده‌ها	شماره فیلد (تگ)	برچسب فیلد
۲۸	#2aIRbIR-503380001$c2020/8/12	801	مبدأ اصلی
۲۹	##$c20070917113926.0130145.0130103.0 $bf-torkashvand	910	اطلاعات مربوط به مستندساز
۳۰	##$af-torkashvand$b20200510103417.0	911	اطلاعات مربوط به ویرایشگر رکورد
۳۱	##$b111673$c1	930	اطلاعات رکورد مستند
۳۲	##bYaa	932	اطلاعات دسترسی به رکورد

همان گونه که جدول ۱ نشان می‌دهد، هر رکورد می‌تواند بیش از صد فیلد فرعی داشته باشد که همه آنها برای تبدیل به اسکاس به‌کار نمی‌رود. این فیلدهای فرعی از نظر متخصصان علم داده، ویژگی محسوب می‌شوند. بررسی دقیق‌تر این رکورد حاکی از وجود داده‌های پرت و داده‌های مفقود در آن است. نمونه‌ای از داده‌های مفقود فیلد فرعی زبان در ردیف‌های ۸، ۹ و ۱۷ است. همچنین در ردیف ۲۲ فیلدهای فرعی، شماره رکورد و زبان و خط ثبت نشده است. داده‌های پرت داده‌هایی هستند که در بعضی مواقع می‌توانند دردسرساز باشند و در بعضی مواقع هم خود مسئله، تشخیص داده‌های پرت و به نوعی تشخیص ناهنجاری[1] است. نمونه‌ای از داده‌های پرت در این رکورد در ردیف ۱۹ قرار دارد که کاربر به جای زبان عربی، زبان عبری را انتخاب کرده است. همچنین < الاشیاء= اشیاء > که در نرم‌افزارهای نسل‌های قبلی به‌کار می‌رفت نیز حاوی علائم پرت است که باید حذف گردد و "الاشیاء= اشیاء" یکی به عنوان اصطلاح گزیده و دیگری در قالب اصطلاح ناگزیده در اسکاس تعریف شود.

مرحله سوم: آماده‌سازی داده

ما به همه داده‌های مبتنی بر مارک اصفا برای تبدیل به اسکاس نیاز نداریم. بنابراین ابتدا باید برای مرحله انتخاب ویژگی‌ها[2] در لایه پیش‌پردازش[3] مهندسی داده[4]، عناصر ضروری برای تبدیل به اسکاس شناسایی و انطباق آنها با فیلدهای مارک تدوین شود. ابزار اسکاس پلی[5] که برای تبدیل داده‌های اصفا در قالب صفحه گسترده اکسل به

1. Anomaly Detection
2. Features Selection
3. Preprocessing
4. Data engineering
5. Skosplay

اسکاس/ آر.دی.اف. به کار می‌رود، الگوهای متعددی دارد که مطالعه آنها به شناسایی عناصر زیر برای تبدیل دادگان اصفا به اسکاس منجر شده است:

- "skos:prefLabel": ویژگی skos:prefLabel این امکان را فراهم می‌کند که یک برچسب واژگانی مرجح (گزیده) را به یک منبع اختصاص دهیم. اسکاس، شکل ساده‌ای از برچسب زدن چندزبانه را امکان پذیر می‌کند. این امر با استفاده از برچسب زبان برای محدود کردن دامنه آن به یک زبان خاص انجام می‌شود (Isaac, & Summers, 2009) در دادگان اصفا، اصطلاحات مرجح به زبان‌های فارسی، انگلیسی، عربی، روسی، تاجیکی به خط سیریلیک و تاجیکی به خط فارسی و فارسی به خط تاجیکی وجود دارد. در مارک ایران، تفکیک این زبان‌ها با ترکیبی از کدهای بین‌المللی زبان و خط امکان پذیر است.
- "skos:altlabel": ویژگی skos:altLabel این امکان را فراهم می‌کند که یک برچسب واژگانی جایگزین را به یک مفهوم اختصاص دهیم؛ برای مثال مترادف‌ها (Isaac, & Summers, 2009). اصطلاحات نامرجح یا ناگزیده در اصطلاحنامه‌های سنتی با این برچسب در اسکاس مشخص می‌شوند.
- skos:broader و skos:narrower بازنمون پیوندهای سلسله‌مراتبی را امکان پذیر می‌کند، مانند رابطه بین یک جنس و نوع، یا بسته به تفاسیر، رابطه بین یک کل و اجزای آن (Isaac, & Summers, 2009). اصطلاحات اعم و اخص در اصطلاحنامه‌های سنتی با این برچسب در اسکاس مشخص می‌شوند.
- skos:related بازنمون پیوندهای همبسته[1](غیرسلسله‌مراتبی) را امکان پذیر می‌کند، مانند رابطه بین یک نوع رویداد و دسته‌ای از موجودیت‌ها که معمولاً در آن شرکت می‌کنند. یکی دیگر از کاربردهای skos:related بین دو دسته است که هیچ کدام عمومی‌تر یا خاص‌تر نیستند. skos:related می‌تواند برای بازنمون پیوندهای کل-جزئی که در قالب روابط سلسله‌مراتبی در نظر گرفته نشده‌اند نیز استفاده شود (Isaac, & Summers, 2009). اصطلاحات مرتبط در اصطلاحنامه‌های سنتی با این برچسب در اسکاس مشخص می‌شوند.
- skos:member: مجموعه‌ها در اسکاس زمانی به کار می‌رود که گروهی از مفاهیم دارای چیز[2] مشترکی هستند، و بهتر است آنها را تحت یک برچسب مشترک گروه‌بندی کنیم، یا جایی که برخی از مفاهیم را بتوان در یک ترتیب معنادار قرار داد. برای مدل‌سازی صحیح چنین ساختارهای مجموعه مفهومی، اسکاس یک کلاس skos:Collection را معرفی می‌کند. نمونه‌های این کلاس،

1. associative
2. Thing

مفاهیم خاص را با استفاده از ویژگی skos:member گروه‌بندی می‌کنند (Isaac, & Summers, 2009). معمولا در اصطلاحنامه‌ها، ساختار درختی این مجموعه‌ها را مشخص می‌کند.

- "URI": در اسکاس و اصطلاحنامه‌های مبتنی بر مفهوم، هر مفهوم با یک شناسگر قراردادی منبع (یو.آر.آی.) شناسایی می‌شود. این شناسگر باید منحصر به فرد باشد (Isaac, & Summers, 2009). از آنجایی که در داده‌های اصفا هر مفهوم یک شماره رکورد دارد، بنابراین از شماره رکوردهای اصفا به عنوان شناسگر یکتا استفاده کردیم.
- skos:note : اسکاس ویژگی skos:note را برای اهداف عمومی سندآرایی ارائه می‌دهد و با الهام از دستورالعمل‌های نظام سازماندهی دانش موجود، برای موارد خاص‌تر از skos:scopeNote، skos:definition، skos:example و skos:historyNote استفاده می‌کند.
- skos:scopeNote برخی اطلاعات (احتمالاً جزئی) را در مورد معنای مورد نظر یک مفهوم ارائه می‌کند؛ به‌ویژه به‌عنوان نشانه‌ای از اینکه چگونه استفاده از یک مفهوم در عمل نمایه‌سازی محدود شده است (Isaac, & Summers, 2009).
- rdf:type:rdf:typeنمونه‌ای از rdf:Propertyاست که برای بیان اینکه یک منبع نمونه‌ای از یک کلاس است استفاده می‌شود. rdf:type(Brickley,Guha,&McBride,2014)برای کلاس مفاهیم concept:skos و برای کلاس مجموعه skos:collection است.

براساس عناصر هسته اسکاس، جدول انطباق بین آنها و فیلدهای مارک ایران تدوین شد (جدول ۲). تجزیه و تحلیل انطباق بین واژگان هسته اسکاس نشان می‌دهد در شناسایی عضویت مفاهیم در حوزه‌ها و اصطلاحنامه‌های خرد اصفا دچار چالش هستیم. در این داده‌ها صرفا در فیلد ۱۵۲$۹ کد حوزه‌های اصلی اصفا ثبت می‌شود؛ به عنوان نمونه برای اقتصاد کد اص_اق. این در حالی است که یک رکورد برای اقتصاد به عنوان مفهوم وجود دارد اما رکوردی برای اقتصاد به عنوان حوزه وجود ندارد. همین شرایط در مورد اصطلاحنامه‌های خرد[1] اقتصاد نیز صادق است؛ مانند اقتصاد بازرگانی، اقتصاد بین‌المللی و ... . بخش عمده‌ای از این مرحله، شناسایی داده‌های کثیف[2] (نادرست) و پالایش و تمیز کردن[3] داده‌ها با حذف یا اصلاح هر خطا یا نبود یکدستی آنهاست. این کار با استفاده از کد ماکرو[4] در اکسل انجام می‌شود که بر کیفیت داده‌ها اثر می‌گذارد. در اکسل، ماکرو کد یک کد برنامه‌نویسی است که به زبان ویژوال بیسیک برای برنامه‌ها نوشته می‌شود. استفاده از کد ماکرو برای

1. Micro-Thesaurus
2. Dirty data
3. Cleaning
4. Macro code

خودکار کردن عملیاتی است که در اکسل دستی انجام می‌شود. این درحالی است که ترکیب کد زبان و خط در اسکاسموس دیده نشده با توجه به نبود امکان تمایز خط اصطلاحات فارسی به خط تاجیکی و اصطلاحات تاجیکی به خط فارسی را می‌توانستیم به ترتیب اصطلاحات ناگزیده زبان تاجیکی و فارسی در اسکاس در نظر بگیریم. برای تفکیک اصطلاحات در زبان‌های مختلف لازم است کد زبان به برچسب skos:prefLabel اضافه شود. اسکاس پلی از مجموعه استانداردهای کدهای زبان ISO 639-1 و مارک ایران از ISO 639-2 استفاده می‌کند. مجموعه کد ISO 639-1 برای استفاده در اصطلاحات، واژگان‌شناسی و زبان‌شناسی ابداع شده و مجموعه کد ISO 639-2 برای استفاده توسط کتابخانه‌ها، خدمات اطلاعاتی و ناشران برای نشان دادن زبان در تبادل اطلاعات، به ویژه در نظام‌های کامپیوتری، ابداع شده است (Byrum, J. D. 1999).

مراحل شناسایی داده‌های پرت و مفقود بر روی عناصر مارک ایران در جدول ۲ انجام پذیرفت. مقادیر مفقود مقداردهی و داده‌های پرت ویرایش شدند. برای مدل‌سازی، داده‌های رده‌ای به هر اصطلاح اضافه گردید. همچنین رکوردهای مجموعه در اصفا شناسایی و در دادگان ایجاد شد.

مرحله چهارم: مدل‌سازی

ویژگی هدف[1] در این پژوهش مربوط به مقدار شناسگری است که برای هر مفهوم باید تولید شود و با استفاده از ساختار سلسله‌مراتبی مفاهیم مشخص گردد هر مفهوم در کدام مجموعه یا اصطلاحنامه خرد قرار می‌گیرد. این مقدار وجود ندارد و از طریق روش خوشه‌بندی سلسله‌مراتبی (از نوع روش با نظارت در چهارچوب کریسپ) با استفاده از ماکروکد در اکسل انجام می‌شود.

مرحله پنجم: ارزیابی

در این مرحله، کیفیت مدل‌سازی –که منجر به تولید مقادیر ویژگی هدف شده است– باید ارزیابی شود. ما باید اطمینان حاصل کنیم که فرآیند شناسایی تعلق هر مفهوم به مجموعه/ اصطلاحنامه خرد هیچ گونه خطا یا تناقضی در داده‌های ما ایجاد نکرده است. برای این کار با توجه به بافت[2] کسب و کار از روش‌های ارزیابی خوشه‌بندی سلسله‌مراتبی، تکنیک بررسی بصری[3] و روش نمونه‌گیری تصادفی[4] استفاده می‌کنیم.

1. Target Feature
2. Context
3. Visual inspection
4. Random sampling

مرحله ششم: توسعه و استقرار

در این مرحله، اصطلاحنامه تبدیل شده خود را در یک محیط مناسب پلتفرم داده‌های پیوندی مستقر می‌کنیم. باید اطمینان حاصل شود که داده‌های ما توسط سایر کاربران و سیستم‌ها قابل دسترس و قابل کشف است. همچنین باید برای کاربرانی که می‌خواهند از داده‌های ما استفاده یا استفاده مجدد کنند، مستندسازی و پشتیبانی ارائه کنیم. برای این منظور، ابتدا داده‌های تمیز ۱۸ حوزه را که مدل‌سازی شده است با ابزار اسکاس پلی به اسکاس در گراف آر.دی. تبدیل کردیم. فایل اسکاس/ آر.دی.اف هرحوزه را با ابزار اعتبارسنجی اسکاس پلی[1] به آزمون گذاشتیم. پس از تایید اعتبار، فایل‌ها را به ووک بنچ[2] انتقال دادیم. ووک بنچ پلتفرمی برای مدیریت هستی‌نگاری‌های اُ.دبلیو.ال.، اصطلاحنامه‌های اسکاس و دادگان عمومی آر.دی.اف. است. این پلتفرم امکان ویرایش چندزبانه اصفا را توسط متخصصان مختلف به زبان‌های مختلف فراهم می‌کند. برای انتشار این دادگان از پلتفرم اسکاسموس استفاده کردیم. در نهایت، داده‌ها در اسکاسموس مستقر و منتشر شد[3] (شکل ۲). این پلتفرم پس از تایید کتابخانه ملی فنلاند در وب سایت اسکاسموس[4] قرار گرفت. اصفا پس از انتشار دادگان خود به تفکیک اصطلاحنامه‌های خرد به *اصطلاحنامه فرهنگی ایران* تغییر نام داد.

شکل ۲- نمایش اصطلاحنامه‌های خرد حوزه اقتصاد اصطلاحنامه فرهنگی ایران در پلتفرم اسکاسموس

1. https://skos-play.sparna.fr/skos-testing-tool/
2. Vocbench
3. http://skosmos.nlai.ir/en/
4. https://skosmos.org/

یافته‌ها

یافته اصلی پژوهش، استقرار و توسعه دادگان اسکاس اصفا در پلتفرم منبع باز اسکاسموس به نشانی skosmos.nlai.ir است. مجموع رکوردها در آغاز پروژه ۱۱۰۰۶ رکورد بود که پس از ایجاد رکوردهای مربوط به حوزه و مجموعه برای خوشه‌بندی به ۱۱۸۸۰ رکورد افزایش یافت. در مرحله آماده‌سازی داده یکی از یافته‌های مهم این پژوهش که در حین اجرای روش شناسی کریسپ به دست آمد، جدول انطباق بین عناصر هسته اسکاس و فیلدهای مارک ایران بود (جدول ۲).

جدول۲- انطباق بین واژگان هسته اسکاس و فیلدهای مارک ایران

ردیف	عنوان	برچسب در اسکاس	فیلد و فیلد فرعی در مارک ایران	برچسب فیلد و فیلد فرعی
۱	شناسگر قراردادی منبع	URI	001	شناسگر رکورد
۲	اصطلاح مرجح به زبان فارسی	skos:prefLabel@fa	250$a	توصیفگر
			$7=fa	کد خط= فارسی- عربی
			$8=per	کد زبان= فارسی
۳	اصطلاح مرجح به زبان انگلیسی	skos:preflable@en	250$a	توصیفگر
			$7=ba	کد خط= لاتین
			$8=eng	کد زبان= انگلیسی
٤	اصطلاح مرجح به زبان عربی	skos:preflable@ar	250$a	توصیفگر
			$7a=fa	کد خط= فارسی-عربی
			$8=ara	کد زبان= عربی

ردیف	عنوان	برچسب در اسکاس	فیلد و فیلد فرعی در مارک ایران	برچسب فیلد و فیلد فرعی
۵	اصطلاح مرجع به زبان روسی	skos:preflable@ru	250$a	توصیفگر
			$7=ca	کد خط= سیریلیک
			$8=rus	کد زبان= روسی
٦	اصطلاح مرجع به زبان تاجیکی	skos:preflable@tg	250$a	توصیفگر
			$7=ca	کد خط= سیریلیک
			$8=tgk	کد زبان= تاجیکی
۷	اصطلاح مرجع به زبان تاجیکی با خط فارسی	skos:preflable@fa	250$a	توصیفگر
			$7=fa	کد خط= فارسی–عربی
			$8=tgk	کد زبان= فارسی
۸	اصطلاح مرجع به زبان فارسی به خط سیریلیک	skos:preflable@fa	250$a	توصیفگر
			$7=ca	کد خط= سیریلیک
			$8=per	کد زبان= فارسی
۹	اصطلاح نامرجع به زبان فارسی	skos:altlabel@fa	450$a,	ارجاع نگاه کنید به
			$7=fa	کد خط= فارسی– عربی
			$8=per	کد زبان= فارسی

ردیف	عنوان	برچسب در اسکاس	فیلد و فیلد فرعی در مارک ایران	برچسب فیلد و فیلد فرعی
۱۰	اصطلاح اعم	skos:broader	550$a	ارجاع نیز نگاه کنید به
			$5=g	اصطلاح اعم
۱۱	اصطلاح اخص	skos:narrower	550$a,	ارجاع نیز نگاه کنید به
			$5=h	اصطلاح اخص
۱۲	اصطلاح وابسته	skos:related	550$a	ارجاع نیز نگاه کنید به
			$5=9	اصطلاح وابسته
۱۳	یادداشت	skos:note	300$a	یادداشت کلی
			330$a	یادداشت دامنه
۱٤	مجموعه	rdf: type skos:Collection	-	–
۱۵	مجموعه	skos:member	-	–

در داده‌های جدول مشاهده می‌شود، در مارک ایران برای مجموعه معادلی وجود ندارد. همان گونه که در بخش روش‌شناسی اشاره کردیم، با بهره‌گیری از ساختار رده‌ای و تکنیک خوشه‌بندی این چالش مرتفع شد.

نتیجه‌گیری

روش‌شناسی به کار رفته در این پژوهش، یکی از روش‌های معمول در پروژه‌های علم داده به نام کریسپ–دی.ام. است. تمام مراحل سه گانه مایلز، راجرز و بکت (۲۰۰٤) در تبدیل اصطلاحنامه‌ها به هستی‌نگاری سبک (کدگذاری آر.دی.اف.؛ بررسی خطا و اعتبارسنجی؛ و انتشار آر.دی.اف.) در گام آخر روش‌شناسی کریسپ–دی.ام. (توسعه واستقرار) قرار می‌گیرد. همچنین مرحله آماده‌سازی روش ون آسم و دیگران (۲۰۰٦) در گام آماده‌سازی داده و تبدیل نحوی؛ تبدیل معنایی؛ و پذیرش استاندارد بین‌المللی اسکاس در گام توسعه و استقرار مدل

کریسپ قابل انجام است. در روش تکامل یافته‌تر باربوسا و دیگران (۲۰۲۱) مراحل انتخاب و استخراج داده؛ انطباق و تبدیل داده‌ها در گام آماده‌سازی و مراحل ارائه آر.دی.اف.؛ سریال‌سازی آر.دی.اف./ تورتل؛ مجوز داده‌های باز؛ و پرس‌وجوی انتشار در گام توسعه و استقرار قرار دارد. نوآوری در این پژوهش جاسازی[1] مراحل روش‌شناسی‌های به کاررفته در ادبیات این پژوهش در چهارچوب کریسپ و نیز اجرای کامل‌روش‌شناسی کریسپ در تبدیل اصفا به هستی‌نگاری سبک اسکاس/آر.دی.اف. است. دادگان میراثی اصطلاحنامه‌ها و سایر نظام‌های سازماندهی دانش، هنگام تبدیل به استانداردهای وب معنایی دچار چالش‌های زیادی هستند که این پژوهش نشان داد این چالش‌ها با رویکرد علم داده، قابلیت حل و فصل دارند. نبود یا کمبود متخصصان کتابداری که بتوانند مهارت کدنویسی ماکرو در اکسل را داشته باشند یکی از چالش‌های مهم این پژوهش بود که بهتر است در پژوهش دیگری از راهکارهای جایگزین بهره مند شویم.

پیشنهادها

- انتشار دادگان پیوندی[2] اصفا: همترازی[3] و برقراری پیوند بین دادگان اصفا با اصطلاحنامه‌های مرتبط در دنیا می‌تواند به انتشار ابرداده‌های پیوندی باز منجر شود که حاوی پیوندها و ارجاع به داده‌های پیوندی باز دیگر است.
- داده‌کاوی و استقرار اصطلاحنامه پزشکی کتابخانه ملی ایران در :skosmos.nlai.ir تکنیک و فرایند داده‌کاوی و استقرار اصفا برای اصطلاحنامه پزشکی کتابخانه ملی ایران نیز قابل تکرار است.
- بهره‌گیری از پلتفرم skosmos.nlai.ir به‌عنوان سرویس متمرکز برای اصطلاحنامه‌ها و هستی‌نگاری‌های میانکنش‌پذیر ایران: اصطلاحنامه‌های متعددی در حوزه‌های مختلف موضوعی در سازمان‌های مختلف تدوین شده است. به منظور حفظ این اصطلاحنامه‌ها در جریان وب معنایی و نیز برقراری پیوند بین آنها و با توجه به ماموریت‌های ذاتی کتابخانه ملی ایران، می‌توان از این پلتفرم برای هم‌افزایی و افزایش هزینه – سودمندی در بودجه کشور بهره‌مند شد.

1. Embedded
2. Linked dataset
3. Alignment

منابع

اکبری داریان، سعیده، انتهایی، علیرضا (۱۳۹۹) (طرح پژوهشی). ارائه مدل پیاده‌سازی اصطلاح‌نامه‌های سازمان اسناد وکتابخانه ملی ایران در چهارچوب‌های وب معنایی SKOS/RDF در محیط نرم‌افزارهای منبع‌باز. سازمان اسناد و کتابخانه ملی ایران.

امیرحسینی، مازیار (۱٤۰۱) (نشست مجازی). سلسله هم‌اندیشی‌های نظام‌های سازمان دانش: سیر تکوین لایه‌های وب معنایی در بررسی جایگاه هستی شناسی‌ها. دانشگاه فردوسی مشهد. https://b2n.ir/Fumlibrary

امیرحسینی، مازیار (۱٤۰۱الف) (نشست مجازی). سلسله هم‌اندیشی‌های نظام‌های سازمان دانش: مهندسی مجدد مفهومی اصطلاحنامه در تدوین طرح مفهومی هستی‌شناسی سبک. دانشگاه فردوسی مشهد. https://b2n.ir/Fumlibrary

References

Akbari-Daryan, Saeedeh, Entehaee, Alireza (2020) (Research project). Implementation of thesauri of National Library and Archives of Iran by Semantic web Frameworks SKOS/RDF in open source applications: present a model. National Library and Archives of Iran. [In Persian]

Amirhosseini, Maziar (1401) (virtual session). The series of common thoughts of knowledge organization systems: The formation process of semantic web layers in studying of ontologies. Mashhad Ferdowsi University. https://b2n.ir/Fumlibrary. [In Persian]

Amirhosseini, Maziar (1401a) (virtual session). The series of common thoughts of knowledge organization systems: the conceptual reengineering of the thesaurus in the development of the conceptual schema of lightweight ontology. Mashhad Ferdowsi University. https://b2n.ir/Fumlibrary [In Persian]

Barbosa, E. R., Dutra, M. L., Godoy Viera, A. F., & Macedo, D. D. J. D. (2021). Thesaurus and subject heading lists as Linked Data.

Transinformação, 33.

Biagetti, M. T. (2021). Ontologies as knowledge organization systems. KO KNOWLEDGE ORGANIZATION, 48(2), 152-176.

Davies, J. (2010). Lightweight ontologies. In Theory and Applications of Ontology: Computer Applications (pp. 197-229). Dordrecht: Springer Netherlands.

Haravu, L. J., & Neelameghan, A. (2003). Text mining and data mining in knowledge organization and discovery: the making of knowledge-based products. *Cataloging & classification quarterly*, 37(1-2), 97-113.

Isaac, A., & Summers, E. (2009). SKOS simple knowledge organization system primer. Working Group Note, W3C.

Martínez-González, M. M., & Alvite-Diez, M. L. (2019). Thesauri and semantic web: discussion of the evolution of thesauri toward their integration with the semantic web. IEEE Access, 7, 153151-153170.

Merriam-Webster. (n.d.). Reengineer. *In Merriam Webster. com dictionary*. Retrieved august 26, 2022, from https://www.merriam-webster.com/dictionary/reengineering

McGraw-Hill (2003). Reengineering. *McGraw-Hill Dictionary of Scientific & Technical Terms,* 6E. Retrieved August 26 2022 from https:// encyclopedia2.thefreedictionary.com/reengineering

McGraw-Hill Companies (2002). reengineering. McGraw-Hill Concise Encyclopedia of Engineering. Retrieved August 26 2022 from https://encyclopedia2.thefreedictionary.com/reengineering

Mazzocchi, F. (2018). Knowledge organization system (KOS): an introductory critical account. Knowledge Organization: KO, 45(1).

Miles, A., Rogers, N., & Beckett, D. (2004). Migrating Thesauri to the Semantic Web-Guidelines and case studies for generating RDF

encodings of existing thesauri. SWAD-Europe project deliverable, 8.

Piatetsky-Shapiro, G (2014). CRISP-DM, still the top methodology for analytics, data mining, or data science projects. https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Theng, Y. L., Foo, S., Goh, D., & Na, J. C. (Eds.). (2009). Handbook of *Research on Digital Libraries: Design, Development, and Impact: Design, Development, and Impact*. IGI Global.

Van Assem, M., Malaisé, V., Miles, A., & Schreiber, G. (2006). A method to convert thesauri to SKOS. In The Semantic Web: Research and Applications: 3rd European Semantic Web Conference, *ESWC 2006 Budva, Montenegro, June 11-14, 2006 Proceedings 3 (pp. 95-109).* Springer Berlin Heidelberg.

Villazón-Terrazas, B. C., Suárez-Figueroa, M., & Gómez-Pérez, A. (2010). A pattern-based method for re-engineering non-ontological resources into ontologies. International Journal on Semantic Web and Information Systems (IJSWIS), 6(4), 27-63.

Zeng, M. L., & Mayr, P. (2019). Knowledge Organization Systems (KOS) in the Semantic Web: a multi-dimensional review. *International Journal on Digital Libraries*, 20(3), 209-230.