

Система мониторинга и сбора статистики для оценки результативности научной деятельности в ОАИ-репозитории JINR Document Server

© Т. Н. Заикина

Объединенный институт ядерных исследований
Дубна

ztanya@jinr.ru

© И. А. Филозова

Irina.Filozova@jinr.ru

Аннотация

В настоящее время результаты деятельности ученых в традиционных научных жанрах активно публикуются в Интернет, размещаются в электронных библиотеках и архивах. Это позволяет осуществлять сбор и производить обработку разнообразной онлайн-статистики. На основе этих показателей можно анализировать распространение влияния отдельных исследований и их авторов.

В 2009 г. в ОИЯИ были начаты работы по созданию репозитория на базе сервера научных документов JINR Document Server (JDS) [2]. Одной из целей создания JDS ставилось обеспечение механизма оценки эффективности научной деятельности сотрудников института.

Для достижения выдвинутых целей было принято решение о разработке новых приложений для JDS с использованием визуализации информации и методов сбора статистических данных, позволяющих повысить эффективность использования и проводить оценку научной деятельности.

1 Введение

Проблема оценки качества деятельности отдельного ученого и научных коллективов появилась с момента зарождения самой науки и всегда являлась одной из актуальнейших и наитруднейших проблем для самой науки и общества. Библиометрические показатели, например, индекс цитирования (citation index), извлекаемые из международных баз данных, являются мощным инструментом при экспертной оценке работы отдельных исследователей и научных коллективов.

В репозитории научных публикаций и документов Объединенного института ядерных исследований (ОИЯИ) JINR Document Server хранится множество атрибутов информационных

объектов, предоставляющих возможность для проведения наукометрического анализа.

2 Система статистики и мониторинга JDS

В результате переноса деятельности исследователей и научных организаций в онлайн-среду создаются необходимые условия для сбора и обработки онлайн-статистики, характеризующей определенные аспекты исследовательских организаций. Практическое использование этой статистики предполагает разработку онлайн-метрики и показателей результатов научно-исследовательской деятельности. Имея такие показатели, можно анализировать распространение влияния отдельных исследований, параметры использования результатов исследований и др.

Несмотря на все недостатки, большинство специалистов сходятся во мнении, что оценка эффективности имеет право на существование. Она должна оценивать реальные результаты работы (а не личные качества сотрудников), вклад служащего в достижение общеорганизационных целей, определять его сильные и слабые стороны, устанавливать, какая профессиональная подготовка ему необходима. Во многих организациях на оценке эффективности построена система оплаты труда.

Одна из главных причин необходимости подсчета оценки эффективности — повышение производительности персонала.

Оценка эффективности позволяет:

- Подвести итог прошлой деятельности и поставить новые цели. Все сотрудники хотят знать, хорошо ли они делают свою работу. Одно из главных преимуществ официальной оценки эффективности заключается в инициировании обратной связи между руководством и сотрудниками, установлении совместных целей на следующий отчетный период.

- Выяснить истинное положение дел. Оценка эффективности помогает удостовериться, что коллектив имеет общие цели и приоритеты.
- Обсудить и скорректировать направления развития.

Таким образом, исследование онлайн-статистики, характеризующей определенные аспекты деятельности ученых и исследовательских организаций, является актуальной задачей.

2.1. Обзор модулей сервера документов ОИЯИ JDS для извлечения статистических данных.

Центральное место в архиве JDS занимают публикации сотрудников ОИЯИ. Но по мере необходимости и пожеланий пользователей в репозитории размещается любая полезная информация, необходимая для поддержки научного исследования, проекта, эксперимента и т. п. ОИЯИ участвует в нескольких крупных международных экспериментах, требующих современной информационной поддержки. В репозитории могут быть размещены коллекции документов, отражающих процесс подготовки, создания экспериментальной установки и получение физических результатов.

CDS Invenio, на основе которой создан сервер документов ОИЯИ JDS, имеет настраиваемый модуль WebStat, обеспечивающий сбор статистических данных о таких параметрах как количество обращений, количество загрузок, наиболее популярные документы, интенсивность поисковых запросов, интенсивность пополнения коллекций и пр. [5]. WebStat позволяет получить статистические отчеты в виде двухмерной графики. Есть возможность оценки документов читателями (модуль WebComment). Таким образом, модуль статистики WebStat совместно с другим инструментарием — модули WebBasket, WebGroup, WebMessage, WebComment — лежит в основе социальной функции сети, позволяет учитывать ее социальные особенности.

Модуль WebComment обеспечивает социально-ориентированной инструмент ранжирования документов читателями и передачи своих замечаний, обмен мнениями по документам. WebMessage реализует кластеризацию (в т.ч. анонимных) пользователей в группы через веб-форумы. Модуль WebBasket позволяет зарегистрированным пользователям создавать, просматривать и управлять личными и групповыми (коллективными) книжными полками. Пользователи JDS имеют возможность взаимодействовать между собой посредством сервиса для обсуждения публикации, образуя, таким образом, сеть коммуникаций в научном сообществе в рамках информационной системы. Анализируя эти связи, можно получать дополнительную важную информацию о характере этих взаимодействий, например, по интенсивности связей между авторами, работающих в одном научном

направлении (теме, проекте) можно судить о степени активности развития данного направления.

2.2. Развитие системы статистики и мониторинга JDS

Статистические данные WebStat доступны только в режиме администратора, но эти показатели должны быть публичны и доступны всем желающим. Поэтому представляется целесообразным иметь систему сбора статистики и мониторинга с развитым набором сервисов для визуализации и графического анализа, что позволит получить более полную картину процессов создания и использования научного знания, текущего состояния и тенденций развития различных направлений разработок и исследований.

Целью проведения мониторинга является обеспечение научного сообщества и лиц, принимающих решение (руководителей групп, секторов, отделов, лабораторий и т.п.) своевременной и актуальной информацией о результатах работы профессионального коллектива.

Использование в совокупности онлайн-овых и библиометрических показателей позволит проводить комплексную оценку уровня исследователей и результатов работы научных коллективов.

Для успешного решения этих задач необходимо дополнить основную статистику (получаемую от модуля WebStat), представить ее в наглядном и компактном виде, т.е. разработать визуальное представление, позволяющее быстро проанализировать состояние информационной системы и эффективно отслеживать протекающие в ее рамках процессы (распределение поисковых запросов по тематическим категориям, интенсивность дискуссий по материалам определенной тематики и т.п.). Т.е. задача может быть решена посредством создания дополнительного инструментария — доступной и удобной системы мониторинга и статистики, функционирующей в реальном времени.

Статистика в ОАИ-архивах, в т.ч. в JDS, разнообразна по содержанию и структуре. В ней присутствует статическая составляющая — набор рассчитанных показателей, таких как коэффициент востребованности публикации, коэффициент востребованности автора, коэффициент влияния публикации и пр., и динамическая — прирост публикаций по годам, распределение публикаций по тематике, распределение авторов по коллаборациям, сети соавторских коллективов и пр. [1]. Статическая информация хорошо представляется в структурированном виде, например, табличной форме. Для динамической информации больше подходит визуальное представление в виде двухмерной графики и анимации.

3 Использование статистической информации и методов визуализации для оценки деятельности авторов JDS

Обработка статистических данных всегда связана с задачей их эффективной визуализации. Ключевая задача визуализации информации заключается в создании понятного пользователю графического отображения набора данных, а также в использовании интерактивных технологий, которые бы упростили работу с данными и позволили пользователю изучать их интуитивно.

Визуализация тысячи объектов с различных точек зрения позволяет увидеть более полную картину и получить больше информации об отображаемых объектах, следовательно, более эффективно исследовать полученную информацию, а также повысить осведомленность пользователя об объектах.

Для выдачи информации конечному пользователю были разработаны 4 шаблона информационных карточек:

- информационная карточка организации (ОИЯИ)
- информационная карточка коллекции
- информационная карточка ресурса (хранимой единицы — препринт, статья и т.д.)
- информационная карточка автора (рис. 1).

В верхней части шаблонов располагаются расчётные значения показателей, в нижней — ссылки на соответствующие визуальные представления (облако ключевых слов, сети соавторских коллективов по различным признакам — принадлежность коллаборации, участие в проектах и т.д.).

Шаблон 1. Информационная карточка организации:

Название	
Количество просмотров	
Количество уникальных просмотров	
Количество загрузок полного текста	
Количество публикаций в зарубежных журналах	
Количество публикаций, выполненных в соавторстве с зарубежными авторами	
Прирост публикаций по годам	
Распределение публикаций по тематике	
Распределение авторов по коллаборациям	
Облако тэгов	
Коэффициенты притягательности	
Сети авторских коллективов	

Шаблон 2. Информационная карточка коллекции:

Название	
Количество просмотров	
Количество уникальных просмотров	

Количество загрузок полного текста	
Прирост публикаций по годам	
Облако тэгов	

Шаблон 3. Информационная карточка ресурса:

Название	
Библиографическая запись	
Количество авторов	
Количество просмотров	
Количество уникальных просмотров	
Количество загрузок полного текста	
Коэффициент востребованности	
Коэффициент влияния	
Цитирование	
Ссылки	
Облако тэгов	

Шаблон 4. Информационная карточка автора:

ФИО	
Принадлежность к организации (affiliation)	
Количество публикаций	
Количество просмотров	
Количество уникальных просмотров	
Количество загрузок полного текста	
Коэффициент востребованности	
Коэффициент влияния	
Публикационная активность	
Облако тэгов	
Данные о цитировании	
Сети соавторских коллективов	

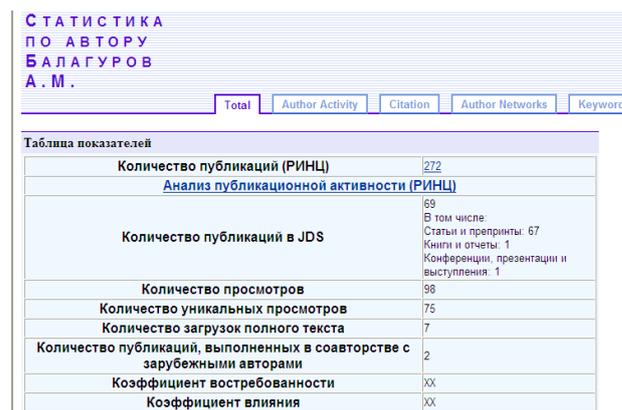


Рис. 1. Пример информационной карточки автора

3.1 Визуализация облака ключевых слов

Облако тэгов (облако меток или облако ключевых слов) стало популярным на многих сайтах и в блогах как визуальное представление списка категорий. Это форма визуализации данных, представляет собой множество ключевых слов, отображенных разными размерами шрифта (или выделенная другим образом). Чем крупнее шрифт,

тем чаще употребляется ключевое слово. Облако тэгов помогает наглядно отобразить, о чем пишет автор, и каков приоритет его интересов. В электронных библиотеках метаданные ресурса включают такую характеристику как ключевые слова.

В системе Invenio модуль BibClassify позволяет автоматически извлекать ключевые слова из полных текстов документов. Автоматическое извлечение ключевых слов для текстовых документов имеет большое значение для обработки информации в электронных библиотеках. Это помогает при каталогизации, классификации и поиске документов. Облако тэгов может быть построено как для отдельного информационного ресурса, так и для коллекции в целом и по совокупности текстов, принадлежащих автору [рис.2].

BibClassify выполняет извлечение ключевых слов на основе повторения специфических терминов, взятых из тезауруса — управляемого словаря всех терминов, имеющих связи в определенном контексте [5].

В системе Invenio используется таксономия High-Energy Physics (HEP), выполненная в синтаксисе RDF / SKOS. Это позволяет получить не только список ключевых слов, но и задать отношения между ключевыми словами и альтернативными способами представления того же ключевого слова.

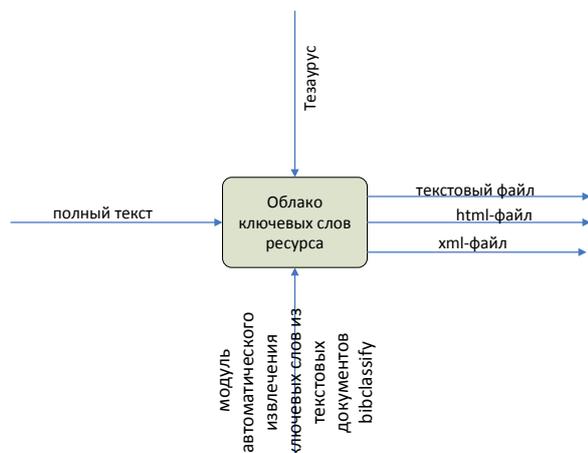


Рис. 2. Формирование облака ключевых слов

Основными элементами SKOS являются [6]:

Концепт (Concept) определяет идею, сущность, объект предметной области;

Семантическое отношение соотносит два концепта друг с другом.

Одиночное ключевое слово Born-Infeld model в таксономии HEP:

```
<Concept
rdf:about="http://cern.ch/thesauri/HEP.rdf#Born-Infeldmodel">
```

```
<prefLabel xml:lang="en">Born-Infeld
model</prefLabel>
<hiddenLabel xml:lang="en">Born-
Infeld</hiddenLabel>
<altLabel xml:lang="en">DBI</altLabel>
<broader
rdf:resource="http://cern.ch/thesauri/HEP
.rdf#fieldtheoreticalmodel"/>
<composite
rdf:resource="http://cern.ch/thesauri/HEP
.rdf#Composite.Born-
Infeldmodelrelativistic"/>
<composite
rdf:resource="http://cern.ch/thesauri/HEP
.rdf#Composite.Born-
Infeldmodelnonlinear"/>
<composite
rdf:resource="http://cern.ch/thesauri/HEP
.rdf#Composite.Born-
Infeldmodelnonabelian"/>
<composite
rdf:resource="http://cern.ch/thesauri/HEP
.rdf#Composite.Born-
Infeldmodelmonopole"/>
<composite
rdf:resource="http://cern.ch/thesauri/HEP
.rdf#Composite.Born-Infeldmodelchiral"/>
</Concept>
```

3.2 Визуализация многопараметрической статистики

Получаемая статистика многопараметрическая, ее трудно изобразить в двухмерном пространстве, поэтому также предлагается использовать динамическую визуализацию. С помощью инструмента Garminder по созданию и публикации интерактивных инфографик были построены прототипы нескольких интерактивных визуализаций, которые предполагается использовать в будущем (рис.3 и рис.4).

Garminder (разработка Google) — удобный инструмент для анализа статистических данных, использующий анимацию. Garminder предлагает удобную подачу и огромные возможности по анализу, позволяя отслеживать показатели в динамике.

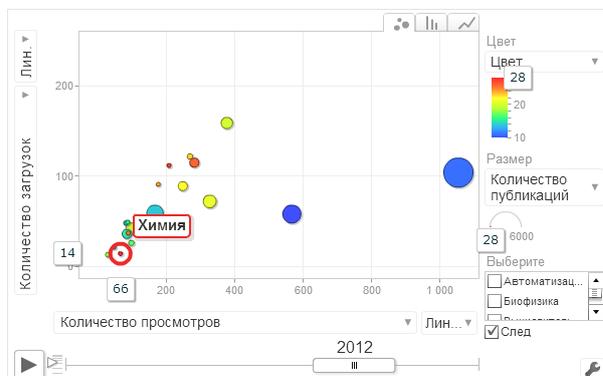


Рис. 3. Динамическая схема пополнения коллекции препринтов ОИЯИ по тематикам.

Пузырьковая диаграмма на рис.3 иллюстрирует динамику изменения роста накоплений публикаций препринтов ОИЯИ по тематикам за время существования архива (2009 – 2013 гг.).

Для отображения данных используются пузырьки разного размера, что позволяет отображать на плоскости динамические процессы. Используется пять переменных: два числовых - по осям X и Y, размер пузырька и цвет, а также переменная времени, которой можно манипулировать с помощью ползунка.

На рис. 3 цветом обозначена тематическая рубрика. Размер пузырька пропорционален количеству публикаций в данной рубрике. По оси X откладывается количество просмотров. По оси Y – количество загрузок.

Пузырьковая диаграмма на рис. 4 иллюстрирует динамику изменения роста накоплений препринтов ОИЯИ по авторам за время существования архива (2009 – 2013 гг.). Для отображения данных используются пузырьки разного размера, что позволяет отображать на плоскости динамические процессы. Используется пять переменных: два числовых - по осям X и Y, размер пузырька и цвет, а также переменная времени.

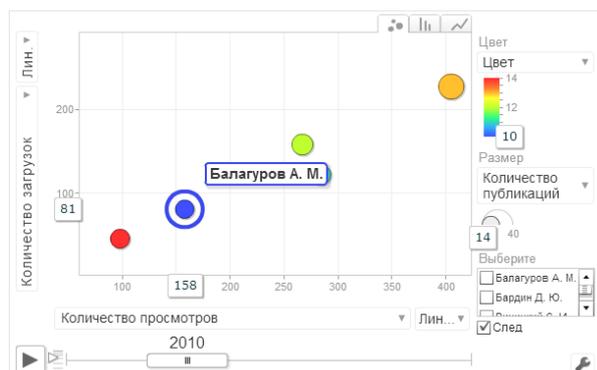


Рис. 4. Динамическая схема депонирования публикаций по авторам.

На рис. 4 цветом обозначен автор. Размер пузырька пропорционален количеству публикаций у данного автора. По оси X откладывается количество просмотров. По оси Y – количество загрузок.

Заключение

Разрабатываемая система статистики и мониторинга в идеале должна отображать в открытом доступе текущие показатели деятельности научного сообщества, обновляемые в реальном времени. Это позволит руководителям и членам научного сообщества получать своевременную и актуальную информацию о результатах работы профессионального коллектива. Используя статистические данные в интеграции с визуализацией, появляется возможность наглядно отобразить зависимости и изменения объектов JDS, позволяя тем самым получить картину научной деятельности авторов и пользователей репозитория.

Литература

- [1] Коголовский М.Р., Паринов С.И. Метрики онлайн-информационных пространств // Экономика и математические методы. – 2008. – Вып. 2
- [2] Куняев С.В. Мусульманбеков Ж.Ж., Устенко П., И.А. Филозова, Шестакова Г. Открытый архив научных публикаций ОИЯИ JINR Document Server.
- [3] Овдей О.М., Проскудина Г.Ю. Обзор инструментов инженерии онтологий // Электронные библиотеки — 2004 — Т. 7, Выпуск 4 — С. 3-19
- [4] Guariano N., Giaretta P. Ontologies and Knowledge Bases. Towards a Terminological Clarification // Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing. Amsterdam: IOS Press, 2009. P. 25–32.
- [5] Invenio-software. <https://Invenio-software.com>.
- [6] Skoseditor. <https://code.google.com/p/skoseditor>.

Monitoring and Statistics System for Impact Evaluation of Scientific Activities of OAI-repository JINR Document Server

T.N. Zaikina, I.A. Filozova

Nowadays, majority of scientific research results published in the Internet, stored in digital libraries and archives, that allows to produce online statistics that characterizing certain aspects of the scientists and research organizations work. On the basis of online indicators we can analyze the spread of influence of individual researches, parameters for the use of research results, etc.

JINR Document Server, JDS, has been launched in 2009 in a framework of Open Access Initiative [2]. One of the JDS goals is to provide mechanism for evaluating the effectiveness of JINR scientific staff.

To achieve the JDS goals, we decided to develop new applications for JDS with using information visualization and statistical techniques for improving using efficiency and evaluate scientific activities.