Proceedings of Symposium in Information and Human Language Technology. Uberlândia, MG, Brazil, October 2–5, 2017. 

# Visualização de glossário em sistemas de recuperação de informação

**Glauber J. Vaz[1], Leandro H. M. de Oliveira[2], Ivo Pierozzi Júnior[1]**

[1]Embrapa Informática Agropecuária
Caixa Postal 6041 CEP: 13083-886 – Campinas, SP – Brasil

[2]Departmento de Pesquisa e Desenvolvimento – Embrapa Sede
Brasília, DF - Brasil

{glauber.vaz, leandro.oliveira, ivo.pierozzi}@embrapa.br

***Abstract.*** *Glossary visualization in information retrieval system is a useful feature for their users. This work presents an information retrieval system component that displays a glossary according to the query results. It also provides details of its implementation and points out various possibilities of development.*

***Resumo.*** *A visualização de um glossário em sistemas de recuperação de informação fornece grande auxílio a seus usuários. Este trabalho apresenta um componente de sistema de recuperação de informação que exibe um glossário de acordo com os resultados das consultas realizadas. Também fornece detalhes de sua implementação e aponta várias possibilidades de evolução.*

## 1. Introdução

Segundo o Dicionário Eletrônico Houaiss da Língua Portuguesa (Versão 1.0 – Dezembro de 2001), na Idade Média e no Renascimento, os glossários representavam a reunião de anotações, antes interlineares (glosas), sobre o sentido de palavras antigas ou obscuras encontradas nos textos e eram apresentados no final de um manuscrito ou até em volumes separados da obra original. Atualmente, não distante do entendimento antigo, os glossários podem ser definidos como um conjunto organizado de termos de uma área de conhecimento e seus significados e definições.

Mais recentemente, os glossários passaram a ser reconhecidos como ferramentas de representação do conhecimento, juntamente com lista de termos e vocabulários controlados, sendo todas já consideradas como Sistemas de Organização do Conhecimento (SOC) [Souza et al. 2010], [Zeng 2008]. Os SOC (mais conhecidos na literatura acadêmica como *Knowledge Organization Systems* ou KOS) têm ganhado cada vez mais atenção em decorrência de sua utilidade em aplicações para a Web Semântica, onde vocabulários formalizados e menos ambíguos, usados como indexadores de

recursos informacionais, são altamente recomendados [Baracho 2016] por causa dos consequentes benefícios que proporcionam no processo de recuperação da informação.

No entanto, vocabulários controlados nem sempre estão comprometidos em apresentar e disponibilizar numa mesma estrutura, seja ela conceitual seja tecnológica, as definições ou as acepções de um determinado termo. Essa funcionalidade é oferecida pelos glossários. Dessa forma, o alinhamento e a convergência dessas ferramentas de representação de conhecimento fornecem uma base bastante consistente para enriquecer sistemas de recuperação da informação, cujas respostas a eventuais buscas podem se beneficiar de maior precisão e menos ambiguidade de um lado e, de outro, podem fornecer ao usuário um panorama terminológico ampliado a ser explorado.

A utilização de um glossário por parte de uma comunidade possibilita uma compreensão homogênea sobre o significado dos termos. A exibição frequente das definições dos termos mais usados por uma comunidade auxilia na consolidação da terminologia e possibilita melhor comunicação entre seus membros, além de criar oportunidades para maior discussão sobre o significado dos termos utilizados. Agregar um glossário a um sistema de recuperação de informação significa oferecer um recurso linguístico justamente no momento em que o usuário está procurando informações associadas ao tema e, portanto, em que está mais aberto a receber esse tipo de informação.

Além disso, no contexto de um sistema de recuperação de informação, a apresentação das definições dos termos contribui nas estratégias de busca do usuário, uma vez que essas definições contêm palavras que podem ser consideradas na formulação de novas consultas por parte do usuário, contribuindo assim no refinamento da busca.

No entanto, não há muitos trabalhos que detalhem a implementação de uma funcionalidade como essa. Um trabalho recente [Bauer et al. 2015] apresenta o WikiHyperGlossary (WHG), tecnologia que usa glossário para possibilitar uma melhor compreensão de documentos da área de Química. Neste sistema, os documentos apresentam termos em destaque com hiperlinks que levam à abertura de novas janelas contendo a definição do termo e outras informações específicas do domínio. No WHG, os documentos são processados por meio de expressões regulares para que sejam identificados os termos relacionados ao tema de interesse. Porém, uma abordagem mais personalizada pode ser utilizada no processamento de documentos, com o uso de diferentes analisadores de texto e não apenas daqueles baseados em expressões regulares. Além disso, no escopo de recuperação de informação, esse processamento pode ocorrer previamente, na fase de indexação, e não apenas durante a geração da marcação HTML.

Neste trabalho, apresentamos um recurso de visualização de glossário em um sistema de recuperação de informação e fornecemos detalhes de sua implementação.

Este recurso é baseado em analisadores de texto personalizados para identificar os termos do glossário presentes nos documentos durante sua indexação. Esta abordagem também cria oportunidades para melhorar a interface de usuário.

## 2. Metodologia

Elasticsearch [Gormley and Tong 2015] é uma tecnologia para a construção de ferramentas de busca. Oferece interface simples via API e é baseada na biblioteca de código aberto Apache Lucene, que oferece recursos de indexação e busca de textos. O processamento de texto é feito tanto na fase de indexação quanto na de busca por analisadores que recebem uma cadeia de caracteres e retornam uma lista de *tokens.* Alguns desses analisadores já são oferecidos pela tecnologia, como, por exemplo, o baseado em expressões regulares, mas outros também podem ser personalizados. São compostos sequencialmente por (i) filtros de caracteres opcionais, que substituem determinados caracteres por outros, (ii) exatamente um *tokenizer*, que produz uma lista de *tokens* a partir de uma cadeia de caracteres, e (iii) por filtros de *tokens*, também opcionais, que podem modificar a lista de *tokens*. A Figura 1 ilustra a composição de um analisador para Elasticsearch. Neste trabalho, explicamos os analisadores usados na implementação do recurso de glossário do sistema.

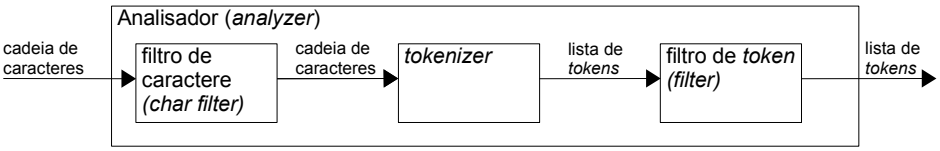


**Figura 1. Elementos de um analisador**

A interface do sistema é baseada em portlets que se comunicam entre si. Os portlets são aplicações que fornecem fragmentos específicos de conteúdo para serem incluídos em uma página de portal [Hepper 2008]. Assim, cada portlet pode ser responsável por uma funcionalidade do sistema de recuperação de informação. Na nossa solução, um portlet foi desenvolvido exclusivamente para expor o glossário.

O modelo de portlet segue o padrão *Model-View-Controller* (MVC). O controlador do portlet faz chamadas à API do servidor Elasticsearch para realizar as consultas. Então, os resultados obtidos são processados para gerar o conteúdo do portlet.

O sistema indexa dados das publicações da Empresa Brasileira de Pesquisa Agropecuária (Embrapa). O glossário utilizado tem 86 termos relacionados a recursos hídricos, mudanças climáticas e agricultura. Sua construção foi baseada na metodologia "OntoMethodus", apresentada e discutida em detalhes por Di Fillipo et al. (2008), sendo realizadas algumas adaptações e complementações ao método. Resumidamente, o itinerário metodológico foi assim executado: (1) construção, limpeza e compilação de um córpus textual sobre o domínio de interesse, composto pela reunião de textos, no

caso, publicações técnico-científicas envolvendo temáticas relacionadas aos impactos ambientais da agricultura e das mudanças climáticas sobre recursos hídricos. O córpus foi composto de textos na língua inglesa e totalizou 1.034.534 palavras; (2) extração semiautomática de candidatos a termos e validação intelectual dessa lista por especialistas nas temáticas mencionadas acima. A validação incluiu a proposição dos termos equivalentes em língua portuguesa, os quais foram obtidos por meio de mapeamento com os dois tesauros agrícolas reconhecidos internacionalmente e que possuem terminologias em português: Agrovoc (http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus) e CAB Thesaurus (http://www.cabi.org/cabthesaurus/). Quando o termo equivalente em português não constava dos tesauros, considerou-se a sugestão dos especialistas, que também foram consultados em relação à seleção de termos a serem definidos, resultando uma lista final de 303 termos; (3) composição de uma base definicional com excertos do córpus textual, ou seja, trechos do texto onde há indicações linguísticas denotativas da definição dos termos selecionados na etapa 2. Nesta fase, a validação dos especialistas também considerou a tradução dos excertos do original em inglês para o português; (4) preenchimento da ficha terminológica (v. modelo na Figura 2) para organização dos dados pertinentes a cada verbete do glossário, constituído por 86 termos. Como etapa final do processo, uma nova validação por especialistas foi realizada visando a melhor adequação do enunciado da definição.

Todo esse itinerário metodológico foi realizado por meio da utilização do software e-Termos (https://www.etermos.cnptia.embrapa.br/index.php), um ambiente computacional colaborativo web de acesso livre e gratuito dedicado à gestão terminológica. O glossário, assim construído, foi salvo em formato JSON e incorporado ao sistema de recuperação de informação.

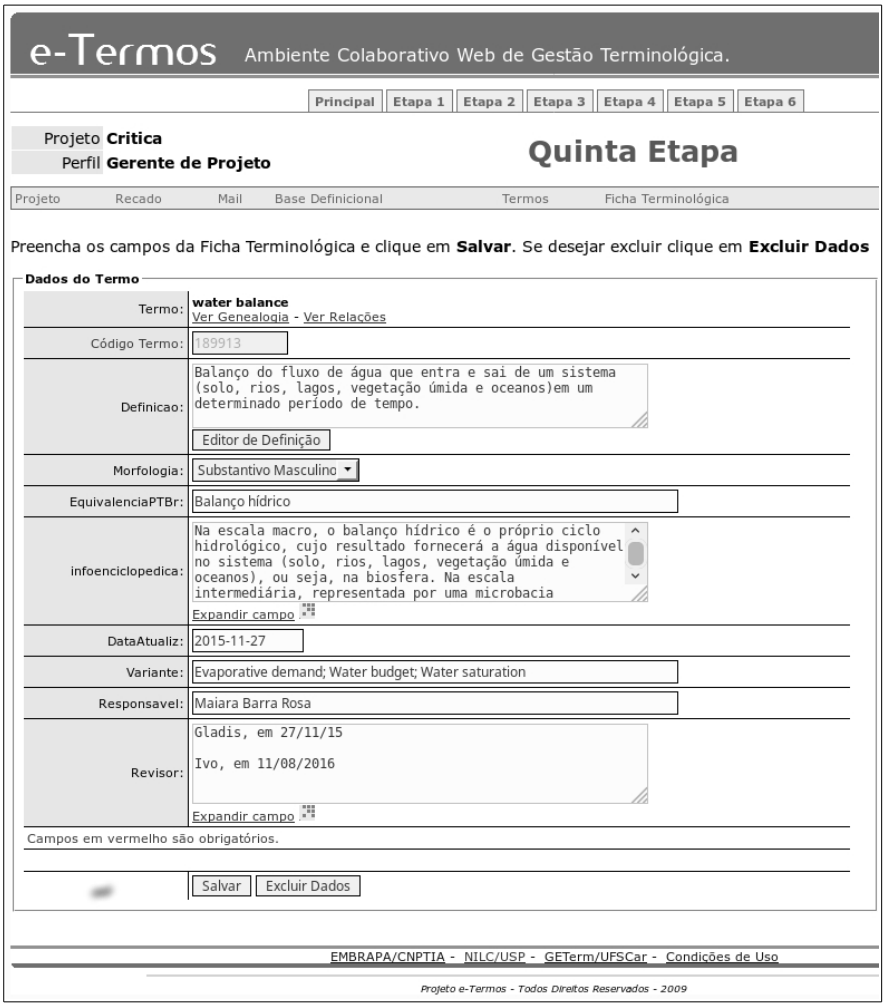


**Figura 2. Ficha terminológica no e-Termos**

## 3. Resultados e Discussão

Desenvolvemos um componente de glossário para um sistema de recuperação de informação que exibe definições de termos presentes em documentos que fazem parte dos resultados de uma consulta. A Figura 3 exibe a interface do sistema com o glossário à esquerda. Este componente foi implementado em um portlet e exibe definições de termos presentes nas publicações listadas nos resultados de uma consulta, exibidos em outro portlet. Os documentos exibidos como resultados podem ser paginados. Assim, são apresentadas apenas as definições dos termos que estão contidos nos documentos exibidos na página corrente dos resultados. Quando uma determinada publicação é selecionada pelo usuário, as definições exibidas são dos termos presentes apenas nesta

publicação.

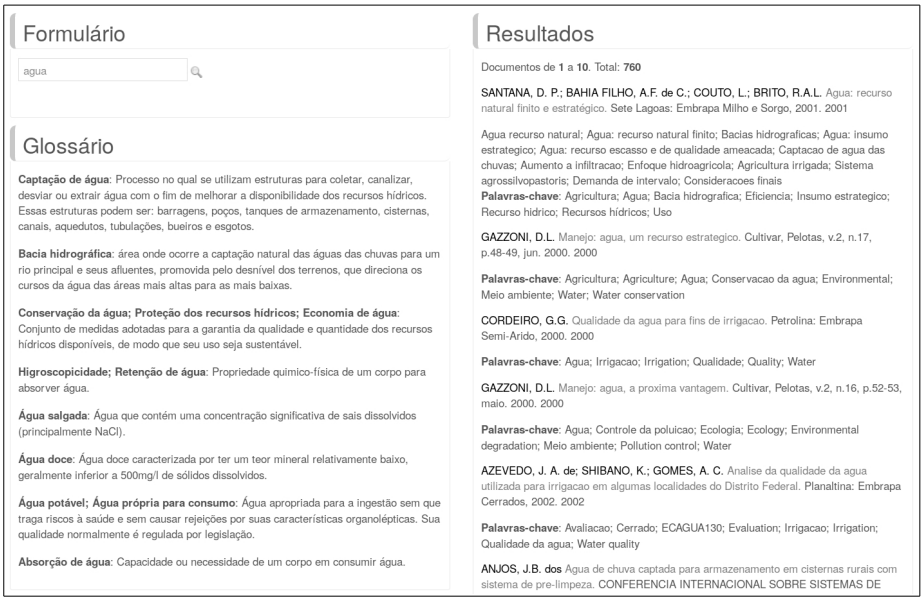


**Figura 3. Exibição do glossário no sistema de recuperação de informação**

Para implementar este componente, dois tipos precisam ser criados no índice utilizado, um para o glossário, em que os termos são indexados e suas definições armazenadas, e outro para as publicações, que inclui um campo que identifica os termos do glossário presentes nas publicações.

A Figura 4 mostra o exemplo de indexação do termo “Bacia hidrográfica”. O tipo “glossario”, portanto, deve conter os campos “TERMO” e “DEFINICAO”. No exemplo, o verbete de “Bacia hidrográfica” é incluído no tipo “glossario” do índice “idx” com identificador “1”. De forma análoga, a Figura 5 mostra o exemplo da inclusão de uma publicação com identificador “10” em “publicacoes” do índice “idx”. Não mostramos aqui todos os campos utilizados no sistema real, mas apenas o suficiente para explicar nossa solução.

```
PUT /idx/glossario/1
{
  "TERMO": "Bacia hidrográfica",
  "DEFINICAO": "Área onde ocorre a captação natural das águas das chuvas para
um rio principal e seus afluentes, promovida pelo desnível dos terrenos, que
direciona os cursos da água das áreas mais altas para as mais baixas."
}
```

**Figura 4: Exemplo de indexação de verbete do glossário**

```
PUT /idx/publicacoes/10
{
  "TITULO" : "Título do artigo",
  "AUTORIA" : "VAZ, G. J.; PIEROZZI JR., I.; OLIVEIRA, L. H. M. de.",
  "FONTE" : "In: STIL 2017",
  "ANO" : 2017,
  "PALAVRAS_CHAVES" : "Linguística; Ciência da Computação",
  "RESUMO" : "Resumo do artigo",
  "TIPO" : "Artigo em Anais de Congresso"
}
```

**Figura 5: Exemplo de indexação de uma publicação**

O mapeamento é o que define como os documentos devem ser armazenados e indexados. No caso do tipo "glossario", o campo "TERMO" deve utilizar analisadores apropriados para indexação e busca de maneira que se mantenha o termo completo em apenas um *token*. Já o campo "DEFINICAO" não é indexado porque não oferecemos buscas diretamente no glossário. Não é disponibilizada, por exemplo, uma funcionalidade que possibilite busca de termos ou definições em função de palavras isoladas presentes nesses campos. O único recurso oferecido por enquanto é a obtenção de definições para termos completos conforme indexados. Por exemplo, considerando-se o caso da Figura 4, uma consulta a "Bacia hidrográfica" obtém sucesso, mas consultas a "Bacia" ou a "chuvas" não.

O mapeamento das publicações é mais complexo. A Figura 6 exibe os trechos relevantes do mapeamento das publicações para o uso do glossário. Trechos ocultos são representados por "…" e ":". Vários campos, ou propriedades, podem ser utilizados para indexar publicações técnicas e científicas. No exemplo, utilizamos os mesmos relacionados na Figura 5 e adicionamos "GLOSSARIO".

Os campos são processados conforme seu tipo e seus analisadores de indexação e de busca. Como estamos tratando principalmente de texto, os tipos são "string" para todos os campos mostrados, exceto para "ANO", tratado como inteiro.

Os valores de determinado campo podem ser copiados para campos complementares, o que é representado pelo uso de "copy_to". Neste caso, copiamos para "GLOSSARIO" todos os campos em que se deseja identificar os termos do glossário: "TITULO", "RESUMO" e "PALAVRAS-CHAVE", pois estão diretamente relacionados ao conteúdo da obra. Campos como nome do autor, por exemplo, não devem ser copiados, porque não se espera encontrar em nomes próprios termos associados ao tema de interesse.

```
PUT /idx/_mapping/publicacoes
"properties": {
        "TITULO": {
            "type":            "string",
            "index_analyzer":  "index_content",
            "search_analyzer": "search_content",
            "copy_to":         [..., "GLOSSARIO"]
        },
        "AUTORIA":{…},
        "FONTE":{…},
        "ANO":{…},
        "PALAVRAS_CHAVES":{…},
        "RESUMO":{…},
        "TIPO":{…},
         :
        "GLOSSARIO":{
            "type":            "string",
            "index_analyzer":  "index_glossary",
            "search_analyzer": "search_glossary",
            "term_vector":     "yes"
        }
},
```

**Figura 6: Mapeamento das publicações**

A análise do campo “GLOSSARIO” é feita por “index_glossary” na fase de indexação e por “search_glossary” na de busca. Estes analisadores devem ser muito parecidos com os utilizados nos campos relacionados ao conteúdo “index_content” e “search_content”. Para garantir um processo de análise equivalente, o “index_glossary” possui a mesma estrutura do “index_content”, mas adiciona um último filtro de *token* em que apenas termos presentes no glossário façam parte da lista de *token* produzida em sua saída. Desta maneira, o campo ‘GLOSSARIO’ de determinado documento contém apenas termos que fazem parte do glossário. O analisador “search_glossary”, por sua vez, pode ser igual ao “search_content” ou também adicionar o mesmo filtro de *token.* De qualquer maneira, a busca encontra apenas termos do glossário no índice.

O campo “GLOSSARIO” ainda utiliza *term vectors*, que armazenam informações adicionais sobre os documentos. Neste caso, para cada documento, é armazenada a lista de termos presentes no texto analisado. Assim, é possível recuperar os termos do glossário que aparecem em determinada publicação a partir de seu identificador.

Portanto, duas operações envolvendo o glossário podem ser realizadas uma vez que seus verbetes e as publicações tenham sido indexados: obter os termos do glossário

presentes em uma determinada publicação e obter as definições dos termos expressos em uma consulta.

Quando no sistema de recuperação de informação o usuário seleciona uma publicação para ver seus detalhes, o controlador do portlet de glossário obtém os termos presentes na publicação e, posteriormente, as definições destes termos. Logo, o conteúdo é gerado para o usuário. Quando várias publicações satisfazem a uma consulta realizada, o controlador do portlet obtém os termos do glossário de cada publicação por meio de chamadas individuais para depois obter as definições de todos estes termos em uma única chamada. Por fim, o conteúdo do portlet é gerado.

O recurso implementado utiliza um glossário cuidadosamente construído por profissionais da terminologia, o que garante a confiabilidade das informações utilizadas. A base indexada no sistema descrito neste trabalho inclui muitas áreas da agricultura, não apenas documentos relacionados à temática do glossário. Além disso, os documentos são publicados predominantemente em língua portuguesa, enquanto o glossário foi baseado em um córpus formado por textos e termos em inglês, com posterior tradução dos termos para português. Ainda assim, dos mais de 80 mil documentos indexados, cerca de 11% apresentaram pelo menos um termo do glossário.

O portlet apenas exibe o glossário, mas pode ser desenvolvido para agregar outras funcionalidades e viabilizar novas formas de interação. *Links* nos próprios termos podem disparar uma nova consulta ou a execução de um filtro nos resultados já obtidos. A oferta de espaço para comentários sobre as definições estabelecidas podem auxiliar na sua melhor elaboração, provocando debates construtivos. Avaliações das definições provêm uma forma de mensurar sua aceitação por parte da comunidade de usuários. Enfim, há inúmeras possibilidades de desenvolvimento para este recurso em um sistema de recuperação de informação.

Tanto a ferramenta apresentada neste trabalho quanto o WikiHyperGlossary são extensíveis, devido ao uso de portlets e APIs. Porém, a análise de texto em nossa solução foi customizada de acordo com as necessidades da aplicação e não se restringiu a expressões regulares, o que pode ser bastante limitante. Além disso, os termos do glossário já são identificados na fase de indexação dos documentos, o que possibilita a apresentação de informações do glossário em conjunto com os resultados de uma busca, sem necessidade de exibição de portlets em novas janelas.

## 4. Conclusão

Neste trabalho, apresentamos um componente que exibe um glossário em um sistema de recuperação de informação, de forma que as definições dos termos sejam exibidas de acordo com os resultados das consultas realizadas. Também fornecemos detalhes de sua implementação, que envolve o uso de portlets para a interface de usuário e de analisadores textuais personalizados que identificam termos do glossário durante a fase

de indexação dos documentos. Esta abordagem, além de abrir oportunidades para o desenvolvimento do recurso de visualização de glossário, possibilita sua integração ao sistema de recuperação de informação de maneira a oferecer grande auxílio a seus usuários.

## Referências

Baracho, R. A. (2016) “Organização e recuperação da informação pilares da arquitetura da informação”. Tendências da Pesquisa Brasileira em Ciência da Informação, v. 9, n. 1.

Bauer, M. A. et al. (2015) "WikiHyperGlossary (WHG): an information literacy technology for chemistry documents", Journal of Cheminformatics, v. 7, n. 1, p. 22.

Di Felippo, A. et al. (2008) “OntoMethodus: a methodology to build domain-specific ontologies and its use in a system to support the generation of terminographic products”. In: Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web, pages 393-395. ACM

Gormley, C. and Tong, Z. (2015). “Elasticsearch: the definitive guide”, https://www.elastic.co/guide/en/elasticsearch/guide/current/index.html, 3 mai. 2017.

Hepper, S. (2008). “JSR 286: Java portlet specification version 2.0”. *Java Community Process*.

Souza, R. R., Tudhope, D. and Almeida, M. B. (2010) “The KOS spectra: A tentative typology of knowledge organization systems”, Advances in Knowledge Organization, v. 12, p. 122-128.

Zeng, M. L. (2008) “Knowledge organization systems (KOS)”, Knowledge Organization, v. 35, n. 2-3, p. 160-182.