

Data Release Interface para o Dark Energy Survey

Luiz A. Nicolaci da Costa^{1,2}, Angelo Fausti Neto^{2,3}, Ricardo L. C. Ogando^{1,2}, Marcio A. G. Maia^{1,2}, Riccardo Campisano^{2,4}

¹Observatório Nacional - MCTI

²Laboratório Interinstitucional de e-Astronomia – LIneA - MCTI

³LSST Corporation - USA

⁴Centro Federal de Educação Tecnológica Celso Suckow da Fonseca – CEFET/RJ

{ldacosta,angelofausti,ogando,maia,riccardo.campisano}@linea.gov.br

Abstract. *Dark Energy Survey, after a proprietary period, releases to community processed data, such as images, catalogs, and maps. LIneA scientific portal allows that these different products, which are input to scientific analysis, are integrated and distributed to the collaboration and public. We describe a query builder to efficiently join these products.*

Resumo. *O Dark Energy Survey, após um período proprietário, disponibiliza para a comunidade seus dados processados na forma de imagens, catálogos, e mapas. O portal científico do LIneA permite que esses diferentes produtos necessários para a análise científica sejam integrados e distribuídos para a colaboração e o público. Descrevemos um query builder para juntar eficientemente esses produtos.*

1. Introdução

O *Sloan Digital Sky Survey* (SDSS) há mais de 15 anos mapeia o Universo estudando de exoplanetas a quasares distantes. Seu sucesso se traduz na publicação de quase 7 mil artigos com a marca SDSS que já foram citados quase 300 mil vezes. Uma das razões do sucesso desse projeto é a disponibilização de novos dados em *Data Releases* (DR) públicos. Imagens, espectros, e catálogos de estrelas e galáxias são disponibilizados através de um site (*SkyServer*) onde astrônomos podem fazer consultas e baixar dados. Desde o DR8, em 2010, o LIneA mantém um espelho desse site¹ que atende a comunidade astronômica brasileira, economizando tempo ao transferir dados e evitando filas de submissão no sítio americano. Além do mais, sob demanda, os pesquisadores brasileiros podem fazer requisições especiais para o uso do site, como por exemplo aumento de quota de espaço em disco, permitindo grandes consultas ao banco de dados do SDSS no Brasil.

O *Dark Energy Survey* (DES) teve sua primeira luz em Setembro de 2012, quando começou o período de Verificação Científica (*Science Verification*, SV), e em Setembro de 2013 começou oficialmente a observação de 1/8 do hemisfério Sul, onde o DES tem condições de repetir o sucesso que o SDSS obteve no Norte. Por exemplo, o DES já revelou a existência de várias vizinhas da Via-Láctea (Drlica-Wagner et al. 2015) e procurou pela contrapartida óptica da primeira fonte de ondas gravitacionais encontrada pelo LIGO (Abbott et al. 2016). Portanto, um *Data Release Interface* (DRI) é uma ferramenta fundamental para o DES e para a comunidade. A seguir discutimos o *Data Server*, protótipo do DRI, já implementado na instância do Portal Científico do DES no Fermilab.

1 <http://www.linea.gov.br/020-data-center/acesso-a-dados-2/>

2. Data Server

O Data Server é um complexo sistema entregue para a colaboração DES em Abril de 2014. Este sistema é composto por um conjunto de ferramentas disponíveis no LIneA e no Fermilab que permitem através de PostgreSQL, TurboGears, e ExtJS:

- Acompanhar o progresso do levantamento DES.
- Verificar visualmente a qualidade das imagens e realizar marcações.
- Sobrepor catálogos às imagens do levantamento.
- Realizar buscas nos catálogos de objetos produzidos a partir destas imagens.
- Criar ou carregar listas de objetos de interesse.
- Visualizar estes objetos e criar “recortes de imagens”

A ferramenta *Tile Viewer* usada na verificação dos dados do DES, é uma das muitas disponíveis no sistema. Nela é possível não só ver imagens, mas associá-las a objetos detectados e, as máscaras aplicadas, como aquelas que indicam estrelas brilhantes. Além disso, associada ao *Upload* de alvos, permite que Grupos de Trabalho verifiquem visualmente a seleção de alvos para observações de seguimento ou aferimento de algoritmos de detecção, sendo intensamente usada, por exemplo, pelo Grupo de Lentes Fortes, cuja detecção de lentes requer confirmação visual (Figura 1).

3. Data Release Interface

Um compromisso assumido com o DES é o desenvolvimento da *Data Release Interface* (DRI) para seu DR público em 2017. Esta interface é uma evolução do *Data Server* discutido acima. Esse processo também é fundamental como preparação para o *Large Synoptic Survey Telescope* (LSST, Abell et al. 2009) cujo volume de dados será ordens de magnitude maior que o do DES. O projeto da DRI, além da interface propriamente dita, envolve a criação de catálogos científicos através da composição de vários produtos, como os de catálogos co-adicionados, de algoritmos de classificação e medição de propriedades, e mapas e máscaras (ex.: estrelas saturadas), mantendo a proveniência dos mesmos (Figura 2). Atualmente, por conta do volume de dados, os catálogos co-adicionados originais, com cerca de 900 parâmetros e centenas de milhões de linhas, são mantidos no Hadoop, pois tentativas com Pgpool encontravam um gargalo ao processar um grande número de regiões do céu em paralelo. Em um *cluster Rocks* com 1000 núcleos, apenas o Hadoop permitiu o processamento de mil regiões do céu em paralelo. Produtos derivados como classificação estrela-galáxia (morfológicos) e *redshifts* fotométricos (ajuste de templates e *machine learning*), menos volumosos, por conterem menos colunas, são ingeridos no banco de dados de catálogos ao longo do processamento usando a ferramenta GAVO DaCHS (Demleitner et al. 2014). Essa ferramenta obedece aos protocolos do Observatório Virtual, melhorando a acessibilidade dos dados no que tange a sua descrição. Finalmente, todas as tabelas são combinadas usando um *pipeline* (*Value Added Catalog*, VAC) cuja componente *query builder* dá ao usuário a capacidade de configurar o catálogo final de modo a se adequar ao seu objetivo científico, seja através de cortes em parâmetros, seleção de regiões do céu, escolha de algoritmos e parâmetros relevantes. Executar todos os passos de processamento, gerando as diferentes máscaras e parâmetros, em um *cluster* com 1.000 núcleos pode levar dezenas de horas, como esperado para esse volume. Mas uma vez prontos, produzir um catálogo para uma região de cerca de 3.000 graus quadrados e cerca de 100 milhões de objetos a partir desses produtos auxiliares leva menos de duas horas. Um dos segredos está em aplicar as máscaras aos catálogos, removendo elementos que não serão utilizados. Finalmente, o *join* apenas entre as tabelas de interesse do cientista provê resultados com eficiência. Esses produtos estão sendo exportados para o banco Oracle do NCSA para a colaboração.

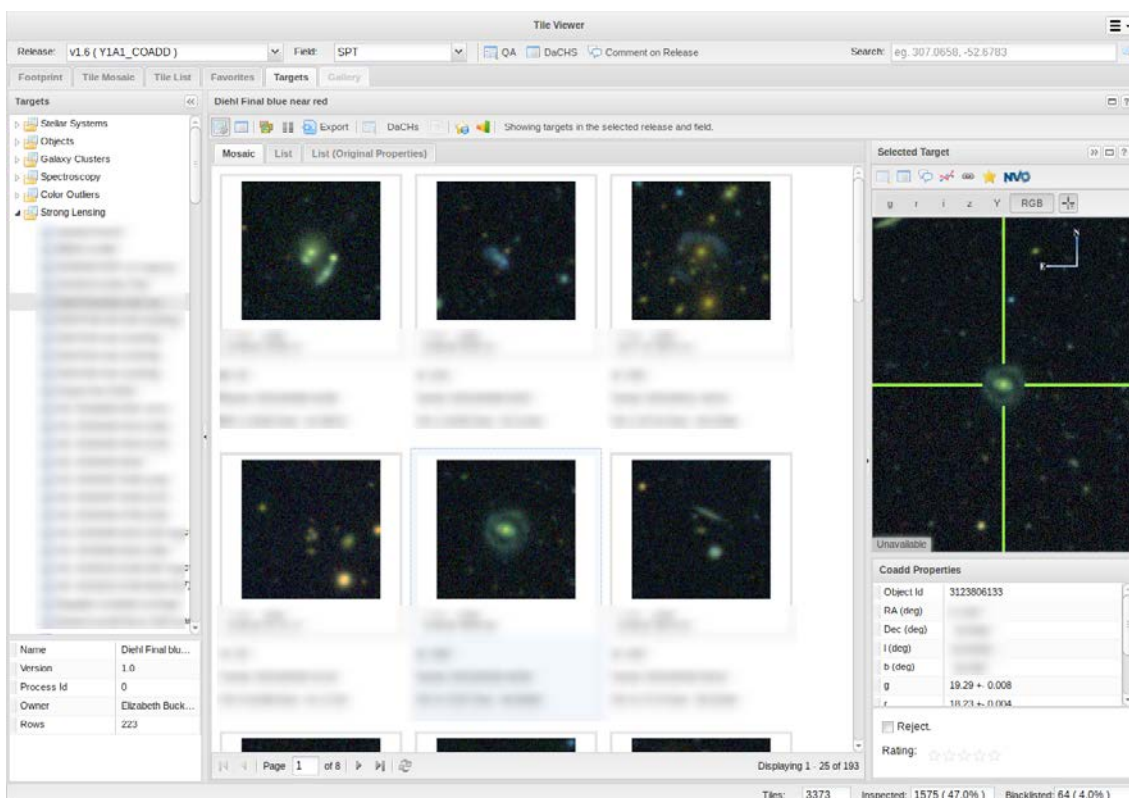


Figura 1. Captura de tela do Tile Viewer no Fermilab mostrando lentes gravitacionais a partir de um catálogo de usuários ingerido no sistema. Os dados das fontes foram encobertos para evitar vazamento.

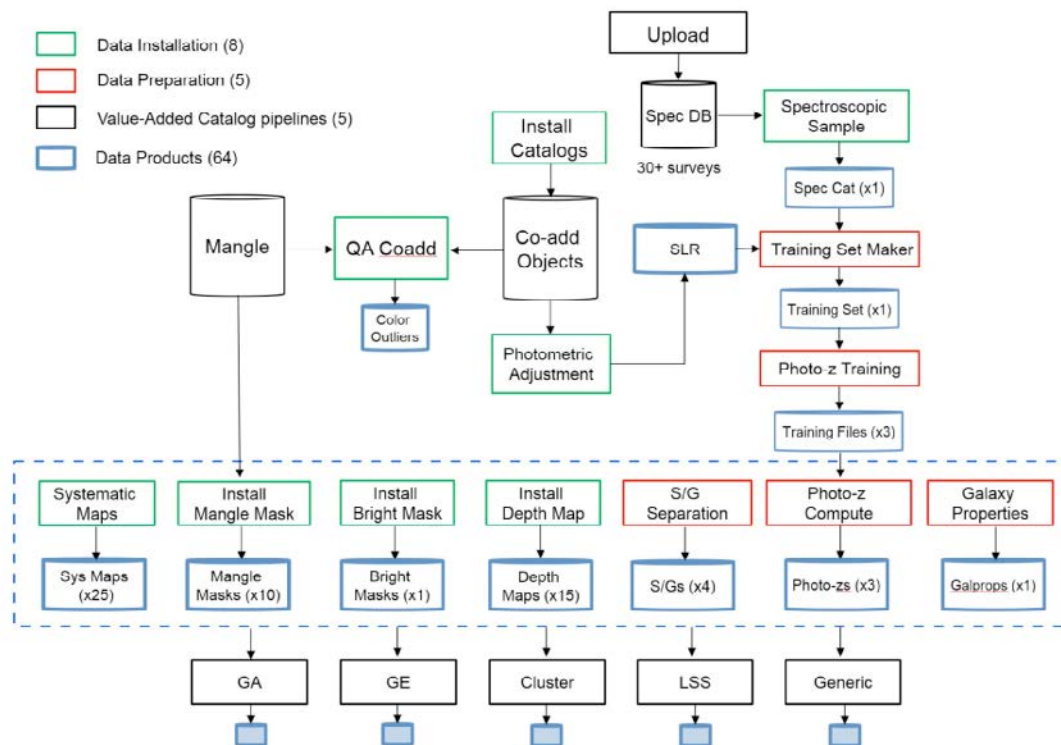


Figura 2. Elementos que compõem os catálogos científicos do DES. Iniciamos com *Install Catalogs* e *Upload* de catálogos espectroscópicos públicos. Os mapas (verde) e os produtos (vermelho) são ingeridos no banco e sua junção produz o VAC.

4. Perspectivas

O sucesso do *Data Server* com a colaboração foi grande e mostrou ao mesmo tempo que ao se disponibilizar um produto complexo como este, é fundamental se preparar para atender a demanda de novos casos de uso. Nos próximos meses, o portal deve continuar seu *refactory* agressivo, de modo a permitir atualizações mais ágeis no OpenStack do NCSA onde o DRI vai operar. Utilizando novas tecnologias como Python 3, Django, e Alladin. Essa interface vai servir inicialmente ao DES e posteriormente a comunidade em geral. Além da importância intrínseca de servir o DES, esse trabalho também serve como ponte para o centro de dados do LSST, o qual vai exigir um esforço ainda maior por parte do LIneA para implementar.

Referências

- Abell et al. (2009) “LSST Science Book, Version 2.0” ArXiv 0912.0201L
- Abbott et al. (2016) “Localization and broadband follow-up of the gravitational-wave transient GW150914” arXiv 160208492
- Alam et al. (2015) “The Eleventh and Twelfth Data Releases of the Sloan Digital Sky Survey: Final Data from SDSS-III” ApJS 219 12
- da Costa et al. (2013) “End-to-end scientific processing in the LIneA Science Portal” CSBC 2013 - BreSci - VII Brazilian e-Science workshop
- Demleitner et al. (2014) “Virtual observatory publishing with DaCHS” A&C 7 27
- Drlica-Wagner et al. (2015) “Eight Ultra-faint Galaxy Candidates Discovered in Year Two of the Dark Energy Survey” ApJ 813 109
- Eisenstein et al. (2015) “The Dark Energy Spectroscopic Instrument (DESI): Science from the DESI Survey” AAS 225 336.05
- Fausti Neto et al. (2013) “Quick Reduce: Dark Energy Survey Camera mountain-top Quality Assessment tool and its master calibration pipeline” CSBC 2013 - BreSci - VII Brazilian e-Science workshop
- Flaugher, B. et al. (2015) “The Dark Energy Camera” AJ 150 150